Scott Soames

THE ANALYTIC TRADITION IN PHILOSOPHY

VOLUME 2

A NEW VISION

The Analytic Tradition in Philosophy

The Analytic Tradition in Philosophy

VOLUME 2
A NEW VISION

•SCOTT SOAMES•

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press, 41 William Street
Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press, 6 Oxford Street
Woodstock, Oxfordshire OX20 1TR

press.princeton.edu

Background pattern courtesy of Shutterstock

ISBN 978-0-691-16003-0

British Library Cataloging-in-Publication Data is available

This book has been composed in Baskerville 10 Pro

Printed on acid-free paper. ∞

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

FOR MARTHA, BRIAN, AND GREG

CONTENTS

PREFACE

This volume continues the story of the early years of the analytic tradition in philosophy told in volume 1. There I chronicled the development of symbolic logic by Frege and Russell, its application to the philosophy of mathematics and the analysis of language, and the efforts by Moore and Russell to refute Absolute Idealism, to beat back American Pragmatism, and to establish a philosophical paradigm based on rigorous conceptual and logical analysis. Although aspects of their emerging paradigm—particularly Russell's logicized version of it—were new, the conception of philosophy it served was not. The aim was to use new analytic means to solve traditional problems of ethics, epistemology, and metaphysics. That changed with the publication of Wittgenstein's *Tractatus Logico-Philosophicus* in 1922, its assimilation by the early Vienna Circle of Schlick, Carnap, and Hahn in the 1920s, and the flowering of logical empiricism in the 1930s. For many philosophers of this new era, analysis wasn't a philosophical tool; it was philosophy. Analysis wasn't (officially) in the service of advancing philosophical theories or developing philosophical worldviews, which, according to the new orthodoxy, must inevitably exceed the limits of intelligibility. Although analysis could be useful in puncturing philosophical illusions, its chief (official) purpose—sketched in the logical empiricists' 1929 proclamation, "The Scientific Conception of the World"—was to formalize, systematize, and unify science. This volume explores the major successes and failures of the philosophers of that era.

Chapter 1 sets the stage by comparing Russell's conception of philosophy in *The Philosophy of Logical Atomism* with Wittgenstein's conception in the *Tractatus*. Although both are versions of logical atomism, the former uses analytic techniques to arrive at a philosophical theory of the world, while the latter uses them to arrive at a philosophical theory of thought and language. Because Russell aimed to explain what reality must be like if our reported knowledge of it is to be genuine, his analyses yielded an analytic metaphysics. Because Wittgenstein aimed to explain what he thought and language must be like if they are to represent reality, his analyses yielded a criterion of intelligibility that proclaimed metaphysics impossible. For Wittgenstein, arriving at this result required explaining

the nature of propositions. To that end, he rejected the Frege-Russell conception of propositions and extracted a new conception from his analysis of meaningful, representational language. My chapters on the *Tractatus* tell this story.

The remainder of chapter 1 explains the abbreviated modal metaphysics with which the *Tractatus* begins. Although it had little impact on later philosophers, and appears to have been written last, it provides the minimal ontological foundation needed for Wittgenstein's conception of propositions. Chapter 2, "The Single Great Problem of the *Tractatus*," explains that conception. Unlike Frege and Russell, Wittgenstein did not take propositions to be the meanings of sentences; instead, he denied that there are such things as sentence meanings. He agreed that propositions are the bearers of truth, but he took them to be something like meaningful sentences, rather than imaginary sentence meanings. On his account, sentences are linguistic facts consisting of expressions standing in syntactic relations. For them to be meaningful is for them to be governed by linguistic conventions. For example, the sentence 'USC is south of UCLA' consists of two names that stand in a relation R—*being followed by the phrase 'is south of', which is followed by.* The conventions governing it stipulate that 'USC' and 'UCLA' are logically proper names of the University of Southern California and the University of California at Los Angeles, and that structures in which two names stand in R are used *to represent* the referent of the first name as *being south of* the referent of the second. One who uses the sentence in this way represents the University of Southern California as *being south of the University of California at Los Angeles*. The truth conditions of the sentence follow from this. In chapter 2, I argue that this analysis of atomic sentences was brilliantly effective. Although Wittgenstein's attempt to extend it to truth-functional compounds and general propositions encountered crippling problems, there is, I argue, a way to solve them.

Chapter 3 examines the idiosyncratic logical system of the *Tractatus,* with special attention to problems arising from its treatment of quantification, identity, and the reduction of metaphysical and epistemic modalities to logical modalities. Chapter 4 focuses on its intelligibility test, the difficulties created for it by the hiddenness of tractarian logical form, the problematic doctrine that one cannot state, in language, the relation between language and the world that allows the former to represent the latter, and the idea that one can show what one can't state. The chapter closes with Wittgenstein's strangely appealing, though questionable, discussions of value, the meaning of life, and the impossibility of philosophy, including the *Tractatus*.

The next seven chapters deal with logical empiricism and contemporaneous advances in logic. Chapter 5 reviews the nineteenth-century scientific positivism of Comte and Mach, along with later work by Hilbert,

Poincaré, Duhem, and Einstein that strongly influenced the Vienna Circle and its allies. Particular attention is paid to Schlick's early theory of knowledge and philosophy of science, which mixed a Kant-style "construction of reality" with a deep appreciation of the new Einsteinian physics. The chapter also explains the effect of a certain natural interpretation of the *Tractatus* in moving leading logical empiricists to a verificationist conception of meaning and, in some cases, a phenomenalistic epistemology.

Chapter 6 examines Carnap's *Aufbau*, intended as a blueprint for unifying all scientific, indeed all objective, knowledge, into a single system. In addition to explaining this goal, and the method for achieving it, the chapter exposes fundamental conceptual difficulties that, in later decades, would prompt improvements in Carnap's position, and help set the agenda for advances in epistemology and philosophy of science. Chapter 7 surveys the triumphalism of logical empiricist writers—Schlick, Carnap, Hahn, Hempel, and Reichenbach—in the first half of the 1930s, whose misplaced confidence in verificationism, the elimination of metaphysics, the philosophical efficacy of the new logic, and the linguistic theory of the a priori tended to overshadow interesting disputes over observation sentences, truth, and the foundations of empirical knowledge.

Chapter 8 is devoted to a cluster of theorems, centered around Gödel's incompleteness results, that revolutionized logic in the 1930s. The aim is to present these theorems in a form that is both accessible to a general reader of analytic philosophy and sufficiently detailed to provide a simple, but moderately sophisticated, understanding of them. After explaining the needed concepts and techniques, I give simple semantic proofs of the Gödel-Tarski theorem of the arithmetical indefinability of arithmetical truth and Gödel's first incompleteness theorem (establishing the incompleteness of formal theories of arithmetic that are true in the intended model). Next, the range of provably incomplete arithmetical theories is extended by proving Gödel's original result—that all *omega-consistent* first-order extensions of a certain weak arithmetical theory are incomplete. I then give Rosser's strengthening of the theorem—that all *consistent* extensions of that weak theory are incomplete. This is followed by an explanation of why second-order arithmetic can be complete without threatening the significance of Gödel's results. His second incompleteness theorem—the unprovability, in certain consistent first-order arithmetical theories, of the consistency of those very theories—results from recreating the reasoning used to prove the first incompleteness theorem within those theories themselves. Finally, the impossibility of a complete, effective procedure for deciding first-order logical truth, or logical consequence, is proven in two ways—one (following Alonzo Church) by reducing it to the incompleteness of first-order theories of arithmetic, and one (following Alan Turing) by reducing it to the halting problem for Turing machines.

Chapter 9 presents Tarski's definition of truth plus his companion definitions of logical truth and logical consequence. In addition to making the technicalities comprehensible, the chapter explains the threat posed by the liar paradox that caused him to seek a definition of truth that was guaranteed not to generate inconsistency when incorporated into metamathematical theories. After explaining his success in achieving this goal, the chapter dissects the illusion that led Tarski, Carnap, and many others to wrongly take his definition to be an *analysis* of truth. By contrast, I argue that his definition of *truth in a model* is a reasonable analysis of the notion of *an interpretation of a sentence* that is needed for a genuine analysis of *logical truth* as truth in all interpretations, and *logical consequence* as truth preservation in all interpretations—provided that the notion of truth in the definition of *truth in a model* is our ordinary one, rather than Tarski's defined notion.

Chapters 10 and 11 deal with two signature doctrines of logical empiricism—the attempted reduction of necessity and apriority to analyticity and the empiricist criterion of meaning. Although both initially seemed reasonable, neither succeeded, for reasons detailed in the chapters. The most interesting failure, perhaps because the doctrine is the hardest to motivate, involves the linguistic theory of the a priori. The key difficulty motivating the theory is traced to a confusion between meaningful sentences and propositions, thought of as uses of such sentences. This confusion—which is sufficient to doom the theory by itself—is essentially the same as the one that had to be resolved in chapter 2 in order to properly reconstruct Wittgenstein's account of atomic propositions.

Part Three of the book investigates contrasting approaches to ethics and metaethics in the 1930s. By far the most influential metaethical view was emotivism, of which Rudolf Carnap, A. J. Ayer, and Charles Stevenson were leading exponents. The arguments for and against this view are discussed in chapter 13, including the now well-known "Frege-Geach problem"—originally advanced in 1939 by W. D. Ross—which ultimately defeated it (without thereby defeating all versions of non-cognitivism). Whereas emotivists rejected the philosophical discipline of normative ethics, other moral philosophers continued to offer normative theories. The most significant new theory of this sort was ethical intuitionism, initiated at Oxford between 1912 and 1930 by H. A. Prichard and most fully developed there by his younger colleague Ross between 1930 and 1939. The considerable strengths, as well as daunting weaknesses, of the views of Prichard and Ross are discussed in chapter 14.

The final view about ethics discussed in Part Three was also the least influential, both during the period and after. I refer to Moritz Schlick's conception of ethics as an empirical science. Unlike Ayer, Carnap, and Stevenson, who sought to *replace* ethics with metaethics, Schlick took metaethics to be preliminary to true ethical theory, which, he believed, was part of empirical psychology. "What?" the modern reader is likely

to exclaim, "Hadn't he heard of the fact-value distinction?" Yes, he had, but he wasn't convinced that one must choose between facts and values in constructing a genuine normative theory. Nor am I, which is one reason why I examine his fascinating book *Problems of Ethics* so closely in chapter 12.[1]

[1] Chapters 5, 6, 7, 8, and 12 of this work are entirely new. Chapters 2 and 3 are almost entirely so. Chapters 10 and 14 are updated and substantially expanded versions of chapters of 12 and 14 of Soames (2003a), volume 1 of *Philosophical Analysis in the Twentieth Century*. Chapters 1 and 4 are updated versions of chapters 9 and 11 of that work, while sections 2, 4, and 5 of chapter 9 are adapted and expanded from Soames (1999), *Understanding Truth*.

Part One

THE *TRACTATUS*

LANGUAGE, MIND, AND WORLD

CHAPTER 1

The Abbreviated Metaphysics of the *Tractatus*

1. AIMS AND SIGNIFICANCE

Volume 1 of this work ended with an extensive discussion of the version of logical atomism found in Bertrand Russell's *The Philosophy of Logical Atomism,* originally presented as eight lectures in 1918. There, we observed Russell's most systematic attempt to use his methods of logical and linguistic analysis, originally deployed in "On Denoting" and *Principia Mathematica,* to craft solutions to what he, along with G. E. Moore, took to be the central problems of philosophy. Moore's own summary of those problems was presented in the first of a series of lectures given in 1910–11 that ultimately were published as *Some Main Problems of Philosophy* in Moore (1953). There, Moore says that the most important, though not the only, job of philosophy is

> to give a general description of the whole Universe, mentioning all the most important things we know to be in it, considering how far it is likely that there are important kinds of things which we do not absolutely *know* to be in it, and also considering the most important ways in which these various kinds of things are related to one another. I will call this, for short, 'Giving a general description of the *whole* Universe', and hence will say that the first and most important problem of philosophy is: To give a general description of the *whole* Universe. [pp. 1–2]

In those lectures, and in the years preceding and following them, Moore showed himself to be highly critical of philosophical descriptions of the universe that *contradicted* what he took to be his commonsense knowledge of it. Included in that knowledge was his knowledge of space and time, past and present, mind and matter, and of other human beings—their material bodies, their conscious states and experiences, and their commonsense knowledge of the same sorts of things that he took himself to know. Although Moore didn't rule out philosophical *additions* to commonsense knowledge, his practice was to subject proposed extensions to relentlessly critical scrutiny—including the Absolute Idealists' arguments for the essential unity and relatedness of all things,[1] J.M.E. McTaggart's vision of human immortality,[2] and William James's insistence on manmade, pragmatic truths.[3] Despite Moore's emphasis on what we know, he did find it puzzling how, exactly, we know all the things we do know. To his disappointment, he never found a satisfying explanation.

Russell was more ambitious. Sharing Moore's traditional conception of philosophy, he employed his method of logical and linguistic analysis to produce a general description of a universe capable of being known without philosophical perplexity. In the years preceding the publication of Wittgenstein's *Tractatus,* the form of analysis Russell used for this purpose in *Our Knowledge of the External World* (1914) and *The Philosophy of Logical Atomism* (1918/19) was the method of *logical construction*. The idea was to arrive at a description of what reality *must* be like, *if* what we take ourselves to know—from both science and ordinary experience– is really capable of being known.

His account of a knowable universe arose from a reductive philosophical analysis of the claims of science and common sense. The aim of the reduction was to show that these claims—which, on their surface, seem to be about entities the existence of which can be known only by philosophically contentious inference—can be interpreted as involving no such questionable entities or inferences. The analysis involved replacing ordinary and scientific claims—the contents of which seem to posit persisting, mind-independent things in "the external world"—with logically complex systems of sentences about epistemically privileged, actual or hypothetical, momentary sensible objects of immediate perception. Just as Russell had earlier attempted to validate our arithmetical knowledge by reducing arithmetical truths to knowably equivalent statements of pure logic—which were (prior to his recognition of the need for the Axiom of Infinity) themselves assumed be transparently knowable—so, in the years immediately preceding the *Tractatus*, he sought to validate our knowledge of

[1] Moore (1919/20).
[2] Moore (1901/2).
[3] Moore (1922).

the external world by reducing statements about it to knowably equivalent, and themselves transparently knowable, statements about perceptual appearances.

In this and succeeding chapters I will present a reading of the *Tractatus* that places Russell's logical-atomist conception of philosophy midway between Moore's traditional conception in *Some Main Problems of Philosophy* and Wittgenstein's radically new conception. In accord with the traditional, but at variance with the tractarian, conception of philosophy, Russell aimed for an all-encompassing theory of the whole universe. In accord with the tractarian, but at variance with the traditional, conception, Russell's official aim was *not* to produce new knowledge of the world unavailable outside of philosophy. On the contrary, the relationship between his system of logical atomism and our pre-philosophical knowledge of the world was meant to parallel the relationship between his logicized version of arithmetic and our pre-philosophical knowledge of arithmetic. Just as his logicist reduction wasn't aimed at giving us new arithmetical knowledge, but rather at validating that knowledge and exhibiting its connections with other mathematical knowledge, so his logical atomism wasn't presented as *adding* to our ordinary and scientific knowledge of the world, but rather as validating it and exhibiting the connections holding among its various parts. It is, at least in part, because Russell thought of his enterprise in this way that he says, in *Our Knowledge of the External World,* that "every philosophical problem, when it is subjected to the necessary analysis and purification, is found to be not really philosophical at all, or else to be, in the sense in which we are using the word, logical."[4]

Russell's view of philosophical problems as essentially logical encompasses the idea that although philosophy has a role to play in describing reality, its task is not to formulate testable hypotheses or to subject them to empirical test. Rather its task is to provide conceptual analyses, which he took to be a kind of creative logical analysis. This is what he had in mind in 1914 when he said:

> [P]hilosophical propositions . . . must be *a priori*. A philosophical proposition must be such as can neither be proved nor disproved by empirical evidence. . . . *[P]hilosophy is the science of the possible.* . . . Philosophy, if what has been said is correct, becomes indistinguishable from logic.[5]

The keys here are the conception of philosophy as a priori and the implicit identification of a priori truths with logical truths, and of a priori connections with logical connections, which it is the task of philosophy to articulate. Since Russell thought that a priori and necessary connections were *logical* connections, he understood the task of revealing and

[4] Russell (1914b), p. 42.
[5] Russell (1914a), at p. 111 of the reprinting in Russell (1917).

explaining them to be a search for philosophically motivated *definitions*, as in the reduction of arithmetic to logic, or *decompositional analyses*, as in his analysis of statements about the external world in terms of statements about perceptible simples.[6] Although the final form of the resulting general description of reality was to come from *philosophical* analysis, the raw material for that general description was seen as coming not from philosophy, but from everyday observation, commonsense knowledge, and empirical science. It was, if you will, an exercise in analytic metaphysics. Russell's atomist system was intended to be an *informative* description of the world, but its informativeness was supposed to lie in our surprise at appreciating what was present all along in the knowledge expressed by the statements of science and everyday life.

This seemingly modest view of philosophy was, in certain respects, not too far from Wittgenstein's more thoroughly deflationary conception of philosophy in the *Tractatus*. However, my statement of Russell's view, which I believe he would have found congenial, is not an entirely accurate statement of his position. As I argued in Volume 1, his "analyses" of ordinary and scientific statements about the world weren't even approximately equivalent to the statements being analyzed. Hence, his resulting atomist system was less an *analysis* of what our pre-philosophical worldview amounts to than it was a proposal to *replace* it with an ambitious and highly revisionary system of metaphysics, driven by an antecedent conviction of what reality *must be like* if it is to be knowable. As we look at the *Tractatus,* we will see that Wittgenstein's thought was not free of its own tension of this general sort—not between what we pretheoretically think the world is like and what it must really be like if it is to be *known*, but between what we pretheoretically think, both about the world and about our own thoughts, and what both the world and our thoughts must really be like if the latter are to represent the former.

If this sounds like the *Tractatus* offers a kind of transcendental metaphysics, there is, I am afraid, no denying that it does. But the tractarian metaphysics is relatively spare, in comparison to the Russellian metaphysics of *The Philosophy of Logical Atomism*, and not intended to be substantively informative in the way that Russell's atomism aspired to be. Although the *Tractatus* begins with abstruse metaphysics, there is no identification of its basic metaphysical simples and virtually no analyses of the statements of science or commonsense. Consequently, there is no attempt to state an informative worldview in which traditional philosophical problems are solved by recasting our ordinary and scientific knowledge into anything purporting to be their true or ultimate form. Rather, the heart of the *Tractatus* is its conception of how thought, which finds its expression in language, represents reality.

[6] See Soames (2014), chapter 12.

Its organizing premise is Wittgenstein's rejection of the conception of propositions found in Frege, the early Russell, and the early Moore, and his replacement of that conception with a new analysis of meaningful, representational language. That Wittgenstein himself saw this as *the single great problem* of philosophy, to be addressed in the *Tractatus*, is suggested by the following passages from the *Notebooks 1914–1916*, which he kept when producing that work.[7]

> My whole task consists in explaining the nature of the proposition. (p. 39)
>
> The problem of negation, of conjunction, of true and false, are only reflections of the one great problem in the variously placed great and small mirrors of philosophy. (p. 40)
>
> Don't get involved in partial problems, but always take flight to where there is a free view over the whole of the *single* great problem. (p. 23)

The single great problem, explaining the nature of the proposition, was, as Wittgenstein then saw it, the problem of explaining meaning, which, in turn, was the problem of finding the essence of representational thought and language. This was both the task of the *Tractatus* and, he believed, the only real task for philosophy.[8]

He took this to be crucial for philosophy because (i) he believed that finding the scope and limits of intelligibility was part and parcel of finding the essence of thought, and (ii) he assumed that in order for a thought (the function of which is to represent the world) to tell us anything intelligible about the world, it must tell us something about which state—among all the possible states the world could conceivably be in—the world really is in. He took it to follow from this that all genuinely intelligible thoughts must be contingent and a posteriori. Since, like Russell, he believed that philosophical propositions are never either contingent or a posteriori, he concluded that there are no genuine philosophical propositions.[9] Since, also like Russell, he believed that all necessary and a priori connections were logical connections, he could, even then, have attempted to offer substantively illuminating logico-linguistic analyses of both scientific and everyday statements, had he shared Russell's belief that the fundamental metaphysical simples that ground all analysis could be informatively identified. But he didn't. On the contrary, he was convinced that it is impossible to informatively identify such objects. Given all this, he had to view his task *not* as solving the traditional problems of philosophy, but as disposing of them.

[7] Wittgenstein (1914–16).

[8] These themes are illuminatingly discussed in chapter 1 of Marie McGinn (2006).

[9] It could be argued that Wittgenstein recognized a single necessary, a priori truth that was empty of content, and so not really representational. But such a vacuous truth could hardly save the conception of philosophy as the search for philosophical truths.

Why then do the first few pages of the *Tractatus* consist of metaphysical pronouncements, which, by the end of the work, are seen as problematic? The mundane, but correct, answer is that Wittgenstein simply saw no way of enunciating, and in his mind establishing, the limits of intelligibility that are the heart of the work without violating those limits in the process. This predicament was not limited to his explicitly metaphysical pronouncements. The *Tractatus* is full of tractarian transgressions. The meager metaphysical sketch with which the work begins was the reflex of his views about how propositions, thought of as (uses of) meaningful sentences of a certain sort, represent the world.[10] His intention was not really to do metaphysics, but to end it by revealing how it violates what is essential to all intelligible, representational thought and language.

2. MODAL METAPHYSICS: FACTS, OBJECTS, AND SIMPLES

1. The world is everything that is the case.
1.1 The world is the totality of facts, not of things.
1.12 The totality of facts determines both what is the case, and also all that is not the case.[11]

What is the case is *what is, or rather what determines what is, true*; while *what is not the case* is *what is, or rather what determines what is, false*. Thus the earliest passages in the *Tractatus* purport to identify the basic elements of reality needed for thought and language to represent it, elements that somehow determine the truth or falsity of all propositions. These elements are identified with atomic facts.

1.13 The facts in logical space are the world.
1.2 The world divides into facts.
1.21 Any one can either be the case or not the case, and everything else remain the same.
2. What is the case, the fact, is the existence of atomic facts.
2.01 An atomic fact is a combination of objects (entities, things).

Here we learn that the facts, the totality of which is the world, are independent of one another, which guarantees that they do not include conjunctive, disjunctive, or negative facts. Rather they must be combinations of objects that somehow suffice to determine which conjunctions,

[10] According to Max Black (1964, p. 27), the initial metaphysical section of the *Tractatus* "was probably the last part to be composed," while being "logically independent" of his "great contributions to philosophical insight" (all of which had to do with logic and language), but "inexorably suggested by Wittgenstein's detailed investigations of the essence of language."
[11] All citations will be from Ludwig Wittgenstein (1922 [1999]), translated by C. K. Ogden. In some cases I will add the Pears-McGuiness translation of the passage (Wittgenstein 1922 [1961]), italicized and in square brackets.

disjunctions, negations, and other complex propositions are true. This, Wittgenstein thinks, is the conceptually minimal way in which we must think of reality, if it is to be represented in our thought and language.

What can be said about the objects that combine to make up atomic facts?

2.02 The object is simple.
2.0201 Every statement about complexes can be analyzed into a statement about their constituent parts, and into those propositions which completely describe the complexes.
2.021 Objects form the substance of the world. Therefore they cannot be compound.
2.0211 If the world had no substance, then whether a proposition had sense would depend on whether another proposition was true.
2.0212 It would then be impossible to form a picture of the world (true or false).

Section 2.02 tells us that there are metaphysically simple objects. These, Wittgenstein will treat as referents of logically proper names. Thus, in a very short space, we are given the ontological counterparts of the two key categories of representational language—proper names and atomic sentences. Section 2.0201 is a compressed statement of his commitment to the fundamental parallel between language and the world. As Wittgenstein will later tell us, an atomic (simple) sentence is a combination of logically proper names that represent the metaphysically simple objects they designate as standing in one or another relation to each other. Thus, sentences are, in effect, structured linguistic entities that are projections of the structured elements of reality they are used to represent. Since all complex sentences are ultimately to be analyzed in terms of the atomic sentences they logically depend on, complex statements are themselves, ultimately, reports about classes of possible atomic facts and the simple objects that make them up. Section 2.021 reminds us that this process of analysis, of moving from the more complex to the less complex, must come to an end—in metaphysically simple objects, on the side of the world, and in logically proper names and atomic sentences composed of them, on the side of language.

So far these doctrines are simply asserted without argument. Sections 2.0211 and 2.0212 are meant to provide an argument for this last claim—i.e., for the claim that the process of decomposition and analysis must terminate in the metaphysically simple. What, precisely, that argument is supposed to be is not made explicit. But given other assumptions of the *Tractatus*, one can make an educated guess. The most likely argument seems to be this: (i) Suppose there were no metaphysical simples. (ii) Then the simplest elements in language—logically proper names—would refer to composite objects; for example, the logically proper name n might refer to an object o, made up of a, b, and c composed in a certain way. (iii) In that case, whether or not o existed, and, hence, whether or not n referred

to anything, would depend on whether or not it was true that a, b, and c were composed in the requisite way. (iv) Since the meaning of n is simply its referent, it would follow that whether or not n had a meaning, and hence whether or not any atomic sentence, or proposition, containing n had a meaning, would depend on the truth of the proposition that a, b, and c are composed in the requisite way. (v) Moreover, if there were no metaphysical simples, then this process could be repeated for a, b, and c—i.e., whether or not it was even meaningful to suppose that a, b, and c were related in the requisite way would depend on the truth of still further propositions—and so on without end. (vi) The process could also be repeated for every name and every atomic sentence. (vii) The result extends to all logically complex sentences, since it is a central doctrine of the *Tractatus* that the meanings of all complex sentences are dependent on the meanings of atomic sentences. (viii) So, if there were no metaphysically simple objects, then whether or not any sentence whatsoever had a meaning would depend on the truth, and hence meaningfulness, of still further statements, the meaningfulness of which would depend on yet further statements, and so on. Since Wittgenstein regarded this scenario as absurd, he concluded that there really must be metaphysically simple objects.[12]

There are two points to notice. First, the argument is based on assumptions about language that Wittgenstein introduces later in the *Tractatus*. Hence, the ontological conclusion he derives here is mandated by his central doctrines about representational thought and language. Second, even if one relies on his linguistic assumptions, one must do more to show that the resulting *reductio ad absurdum* really reaches an absurdity, and so justifies his final conclusion. Why is it absurd that the meaning of some, perhaps even all, sentences should depend on the truth of further propositions?[13]

In answering this question it is crucial to clarify what one means by saying that the meaning of one sentence, P, depends on the truth of another

[12] Taking it to be established that there are metaphysically simple objects, a defender of the *Tractatus* might extend the argument to show that *only metaphysical simples* are constituents of (atomic) facts. For suppose otherwise, i.e., that some object o entering into a possible atomic fact F(o) is not metaphysically simple. Then o is a composite object made up of objects a, b, and c composed in a certain way. Thus, o's existence depends on there being a fact F(a,b,c) of those objects being combined in the required way. Since this violates the independence of facts stated at 1.21, o must not be composite.

[13] It may be noted that while the argument makes crucial use of the notions *sentence* and *proposition*, it doesn't say what propositions are or how they are related to sentences. Roughly put, Wittgenstein took propositions to be something like meaningful sentences, uses of such sentences, sets of sentences that mean the same thing, or, as it is put in Ramsey (1923), abstract types the instances of which are sentences that have the same sense. For now, the key point is that, for Wittgenstein, propositions are closely related to sentences in a way that remains to be made precise. The task of reconstructing a coherent view of this type will be taken up in chapter 2.

sentence, or proposition, Q. Suppose one means that in order to determine, or come to know, that P is meaningful (as well as coming to know what P's meaning is) one must *first* determine, or come to know, that Q is true. On this interpretation, what is said in the argument to be absurd is that in order to determine, or come to know, that any sentence has a meaning (as well as to know what it does, in fact, mean), one has *first* to determine, or come to know, that other sentences are both true and meaningful, and so on, *ad infinitum*. That really is absurd, since it leads to the result that we can never determine, or come to know, what any sentence means, or whether it was meaningful at all.

But the argument doesn't establish that this absurdity follows from the supposition that there are no metaphysical simples, since, on this interpretation, steps (iii) and (iv) do not follow obviously from step (ii). To see this, suppose I were to use the word 'this' as a logically proper name to refer to the chair I am sitting on. In order for this use of the word to have that meaning, the chair I intend to use it to refer to must exist. Suppose that my chair is made up of a huge collection of molecules configured in a certain way. Since my chair is made up of these molecules in this configuration, it may be necessary in order for my chair to exist, and, hence, in order for my use of the word 'this' on the present occasion to have both a referent and a meaning, that these molecules be so configured.[14] But this is *not* something I have to know in order to know that the chair exists, or that my utterance meant what I took it to mean.

Next imagine a group of people with no conception of molecular structure who speak a language L with precisely the logical structure that Wittgenstein imagines, where the logically proper names are restricted to referring to people and ordinary middle-sized objects of their acquaintance. Even if none of the names, atomic sentences, or non-atomic sentences of L would have meanings were it not for the fact that certain molecular configurations existed, speakers of L could know their words to have the meanings they do without knowing any of this. The reconstructed tractarian argument for metaphysical simples fails because it doesn't, as it stands, rule out the possibility that our language might be like L in never referring to metaphysical simples.

One could, of course, repair it so that steps (iii) and (iv) really did follow from step (ii). For example, one could stipulate that for the meaningfulness of a sentence S to *depend* on the truth of the claim that so-and-so is simply for it to be the case that *necessarily*, *were it not a fact that so-and-so,*

[14] Perhaps not every molecule of which my chair consists must exist in order for my chair to exist, and perhaps the relevant cloud of molecules need not be configured exactly as they are presently configured. Still it may be necessary that if my chair exists, then some large number of the relevant molecules must be configured in some way approximating how they are presently configured in order for my chair to exist. The argument abstracts away from fine details of this sort.

then S would not be meaningful (or at least have the meaning it does). But, with this interpretation of dependence, the conclusion derived from the supposition that there are no metaphysical simples is no longer obviously absurd. Why shouldn't it be the case that for any sentence S, S wouldn't have a meaning (or at any rate have the meaning it does) were it not a fact that so-and-so, which, in turn, would not have been a fact had not it also been a fact that such-and-such, and so on, *ad infinitum*? Perhaps there is some good reason for thinking that this really is impossible, or absurd, but, if so, we haven't located it.

So far we have two versions of the argument. One rests on a claim about what knowledge of meaning epistemically requires; the other rests on a claim about what having a given meaning metaphysically requires. As we have seen, the former version is, though a genuine *reductio*, unsound, while the latter is no *reductio*. There is, however, a tractarian premise that could be added to bring these two versions together in a way that might more plausibly be thought to establish Wittgenstein's conclusion. The needed tractarian premise relates necessity to apriority, and ultimately to provable logical truth. The premise, which will be discussed in later chapters, is that *a proposition is necessarily true if and only if it is knowable a priori, if and only if it is a logical truth that can be proven by formal calculation*. Although I take this to be one of the central philosophical errors of the *Tractatus,* Wittgenstein and his followers took it to be an important truth.

With this in mind, consider again the hypothesis that o is a composite object that consists in objects a, b, and c combined in a certain way. Given this, one might be able to argue that it is a *necessary truth* that o exists if and only if a, b, and c are combined in the right way.[15] It then follows from the tractarian collapse of metaphysical, epistemic, and logical modalities into one another that it is *knowable a priori* that if o exists, then a, b, and c are combined in such-and-such way. But then, the proposition *that a, b, and c are combined in such-and-such way* must be an *a priori consequence* of the proposition *that o exists*. Next it is argued that no agent who is not in a position to know *that a, b, and c are combined in such-and-such way* can know *that o exists*. Now return to the example about the chair I am sitting on and the complicated configuration of molecules with which it is identified. I don't, in fact, know which molecules are present in the array, or how they are related to one another. Moreover, there is no way for me to derive the correct conclusions about this from the proposition that I express by saying "This chair exists." Since I am not in a position to know that the molecules (my a, b, and c) are combined in the requisite way, it follows that I don't know *that this chair—o—exists* after all.

I don't accept this conclusion, because I take the tractarian collapse of the modalities on which it is based to be a mistake. But logical atomists

[15] The qualifications mentioned in the previous footnote apply here as well.

like Russell and Wittgenstein couldn't avoid the conclusion in this way. Suitably interpreted, they wouldn't reject it at all. The Russell of *The Philosophy of Logical Atomism* would express the conclusion by saying that my chair is a *logical fiction,* meaning by this that although the sentence 'the chair SS is sitting on exists' is true, a proper analysis will reveal that it *doesn't* assert the existence of any entity properly characterized as a chair or as something I am sitting on.[16] A proper analysis must reveal this if, as Russell and Wittgenstein believed, all necessary, conceptual connections between propositions are nothing more than logical connections to be made transparent through analysis. Applying this idea to the sentence about my chair, they would claim that it speaks of metaphysical simples (which chairs are obviously not) as being arranged in a certain way, and nothing more. For Wittgenstein, there are no composite objects because if there were, they could be named by logically proper names, with the result that some necessary connections between propositions wouldn't be logical or a priori connections.[17] He would say that the fact that I do know the truth expressed by 'the chair SS is sitting on exists' without knowing anything about molecules just shows that molecules aren't simples. If we could informatively identify the simples, we could specify just what simples we are talking about, and what we are saying about them. But, as we are about to see, it is central to the *Tractatus* that we can't do this.

Putting this all together, we can improve the reconstructed tractarian argument for metaphysical simples as follows. (i) Suppose there were no metaphysical simples. (ii) Then the simplest elements in language—logically proper names—would refer to composite objects; for example, a logically proper name n might refer to an object o, made up of a, b, and c composed in a certain way. (iii) In that case, it would be both a necessary and a priori truth that n exists iff a, b, and c are composed in the requisite way. (iva) Since the meaning of n is simply its referent, it would follow that *knowing* that n means what it does, and hence knowing the meanings of atomic sentences containing n (and perhaps even knowing that they are meaningful) would require *knowing* the proposition that a, b, and c are composed in the right way. (ivb) Because tractarian propositions are meaningful uses of sentences, this would, in turn, require having proper names a*, b*, and c* for a, b, and c, and using them in a proposition—*that a, b, and c are indeed combined*—that one knows to be true. (v) Moreover, if there were no metaphysical simples, then this process could be repeated for a, b, and c—i.e., *knowing* that they exist and that propositions about them are meaningful, and have the senses that they do, would require *knowing*

[16] For critical discussion of Russell on analysis and logical fictions, see pp. 614–29 of volume 1 of this work.

[17] If n named a composite object, then, since names are rigid designators, it would be plausible to suppose that ⌜n exists only if so-and-so are combined in such-and-such way⌝ is a necessary truth that can't be known a priori, and certainly is not a logical truth.

the existence of still further objects, as well as the meaningfulness of still further names for those objects and the truth of atomic propositions about how they are combined—and so on without end. (vi) The process could be repeated for every name and every atomic sentence. (vii) Finally, the result extends to all logically complex sentences, since it is a central doctrine of the *Tractatus* that the meanings of all complex sentences depend on the meanings of atomic sentences. (viii) Thus, if there were no metaphysically simple objects, then one couldn't *know* the meaning of any sentence, or perhaps whether it even had a meaning. Since unknowable meanings are not meanings, the supposition that there are no metaphysical simples leads, in the presence of other tractarian assumptions, to the absurd conclusion that no sentences are meaningful. This is Wittgenstein's *reductio*.

This is not the place to critique the cogency of the various tractarian assumptions on which the argument depends. For now it is enough to emphasize that the notorious tractarian collapse of the modalities was one of the key doctrines at work in motivating the simplicity of objects, which was fundamental to the ontology of the *Tractatus*.[18] The resulting picture involves a striking parallel between language and reality. Linguistically simple expressions (logically proper names) stand for ultimate metaphysical simples. Linguistically simple sentences, which are combinations of names standing in relations to one another, stand for atomic facts, which are combinations of metaphysical simples standing in relations to one another. Since complex sentences will be claimed to be truth functions of atomic sentences, a world of atomic facts is all that is needed to determine the truth of all meaningful sentences. Whether the ontology is really derived from the linguistic theses, or whether each plays a role in motivating the other, the two are designed to fit together as hand and glove. The resulting metaphysical vision is a sparse but logicized version of traditional metaphysical atomism.[19]

[18] It is instructive to compare the above reconstruction of the *reductio* with Max Black's summary account of it on page 60 of Black (1964). He says, "'If the world had no substance'—i.e. if there were no [simple] objects— . . . we could never know what the sense of a given S_1 was without first, *per impossibile*, knowing an infinity of other propositions to be true. More simply: unless *some* signs are in direct connection with the world (as names are when they stand for [simple] objects) no signs can be in indirect connection either. Thus the sense which we find attached to the propositions we encounter in ordinary life forces us to believe in elementary propositions and so to believe in [simple] objects." On p. 57 Black notes that the regress is closely connected to the possibility of *analysis* in Wittgenstein's sense, and on p. 59 he generates the regress by taking composite objects to be *necessarily* composed of their parts. What he doesn't do is explain why or how this necessity imposes requirements about what must be *known* by one who knows that the object exists, and hence that a name for it is meaningful. This is the role of the tractarian thesis collapsing the modalities, which drove both the logical atomism of Wittgenstein and that of Russell. I suspect the reason that Black didn't mention it is that it was not, in 1964, recognized to be the very far-reaching and deeply misguided thesis that it is.

[19] This point is emphasized in Robert Fogelin (1987).

3. WITTGENSTEIN'S LOGICALLY ATOMISTIC EXPLANATION OF CHANGE AND POSSIBILITY

Traditional atomism held that there are certain simple, indivisible bits of matter called 'atoms' which are the building blocks out of which everything in the universe is made up. All change in the universe was held to be the result of old combinations of atoms breaking down and new combinations taking their place. Even though atoms were taken to be the source of all change, they were themselves regarded to be eternal and unchanging.

Wittgenstein took over this traditional picture and recast it in a new form. The traditional statements of atomism looked like very general empirical hypotheses that might eventually be confirmed, refuted, partially supported, or partially undermined by continuing progress in science. Wittgenstein's version of atomism was different. His statements couldn't be confirmed or refuted by science because they were supposed to be prior to science. In addition, the simples he talked about were not simply the unchanging source of all change; they were also the source of all conceptual or logical possibility. Just as all change, all variation over time, is the combination and recombination of unchanging simples, so all variation in *logical space* between one possible state of affairs and another is a matter of the way that the same metaphysical simples are combined.

Wittgenstein expresses this idea in various ways. For example, in sections 2.027, 2.0271, and 2.0272 we get the idea that metaphysically simple objects are the unchanging source of all change.

> 2.027 The fixed, the existent and the object are one. [*Objects, the unalterable, and the subsistent are one and the same.*]
>
> 2.0271 The object is the fixed, the existent; the configuration is the changing, the variable. [*Objects are unalterable and subsistent. Their configuration is changing and unstable.*]
>
> 2.0272 The configuration of the objects forms the atomic fact.

Wittgenstein also makes it clear that the metaphysically simple objects of the world exist at all possible states of the world, and are the source of all possibility. On this view, to say that something isn't the case, but could have been, is to say that although the basic objects are not combined in a certain way, they could have been so combined. Sample passages indicating this view include the following.

> 2 What is the case, the fact, is the existence of atomic facts. [*What is the case—a fact—is the existence of states of affairs.*]
>
> 2.01 An atomic fact is a combination of objects (entities, things). [*A state of affairs (a state of things) is a combination of objects (things).*]
>
> 2.011 It is essential to a thing that it can be a constituent part of an atomic fact. [*It is essential to things that they should be possible constituents of states of affairs.*]

2.012 In logic nothing is accidental: if a thing *can* occur in an atomic fact the possibility of that atomic fact must already be prejudged in the thing. [*In logic nothing is accidental; if a thing* can *occur in state of affairs, the possibility of the state of affairs must be written into the thing itself.*]

2.0121 (c) A logical entity cannot be merely possible. Logic treats every possibility and all possibilities are its facts. [*Nothing in the province of logic can be merely possible. Logic deals with every possibility and all possibilities are logical possibilities.*]

2.0122 The thing is independent, in so far as it can occur in all *possible* circumstances, but this form of independence is a form of connection with the atomic fact, a form of dependence. . . . [*Things are independent in so far as they can occur in all* possible *situations, but this form of independence is a form of connection with states of affairs, a form of dependence. . . .*]

2.0123 If I know an object, then I also know all the possibilities of its occurrence in atomic facts. [*If I know an object, I also know all its possible occurrence in states of affairs.*]
(Every such possibility must lie in the nature of the object.)

2.0124 If all objects are given, then thereby are all *possible* atomic facts also given. [*If all objects are given, then at the same time all* possible *states of affairs are also given.*]

2.014 Objects contain the possibility of all states of affairs.

2.0141 The possibility of its occurrence in atomic facts is the form of the object.

2.021 Objects form the substance of the world. . . .

2.022 It is clear that however different from the real one an imagined world may be, it must have something—a form—in common with the real world.

2.023 This fixed form consists of the objects.

According to the *Tractatus*, simple objects are fixed and unchanging. All possibility and all change are understood in terms of the combinations and recombinations of the same simple objects. Clearly, the individual simples persist throughout time, and exist at different possible world-states. There are strong suggestions that they exist throughout all time and at every possible world-state. In the *Tractatus*, all possibility—all variation in logical space—is nothing more than variation in the way that metaphysical simples are combined. But what are these objects like? From what we have said so far, one might think that they are something like the tiny billiard-ball bits of matter envisioned in traditional versions of atomism. But this isn't what Wittgenstein had in mind.

4. THE HIDDENNESS OF THE METAPHYSICALLY SIMPLE

Wittgenstein says that objects are simple. They are shapeless, colorless, and, in general, have none of the familiar properties exemplified by

ordinary medium-sized things we encounter in everyday life. Not only do metaphysical simples lack those familiar properties; they are what, so to speak, make up or constitute such properties. One might say that the familiar properties of everyday life "come into existence" only with the configuration of simple objects. For this reason, we have no way of describing such objects, though, supposedly, we can name them.

Wittgenstein makes an illuminating comment about shape in the notebooks he kept while working on the *Tractatus*. He says:

> Let us suppose we were to see a circular patch: is the circular form its property? Certainly not. It seems to be a structural "property". And if I notice that a spot is round, am I not noticing an infinitely complicated structural property?[20]

The point is something like this: when we say that something we perceive is circular, what we are really saying is that the metaphysically simple objects that make it up bear certain structural (in this case, spatial) relations to one another. Thus, the logical form of a sentence *the so-and-so is circular* is, or at least includes, a complex statement of the sort *a is related to b in such-and-such way, which in turn is related to c in a certain way, which in turn is related to d* (and so on). Here 'a', 'b', 'c', and 'd' are logically proper names for metaphysical simples that make up the complex thing denoted by the subject of the original sentence. On this view, all talk of circularity can be analyzed into talk of how multitudes of simples are related to one another. If we ask whether the metaphysical simples are themselves circular, we are asking a nonsensical question. To say that something is circular, or that it has any shape, is to presuppose that it is a complex, the parts of which stand in relations to one another. Since, by definition, simples have no parts, they have no shape.

What applies to shape also applies to other familiar properties encountered in everyday life. Whenever we say of anything that it has one of these properties, what we are saying is that the simples that make it up are arranged in a certain way. Since all these properties arise only at the level of combinations of simples, it is nonsensical to ascribe them to the simples themselves. We can, in principle, name the simples with logically proper names, and say something about how they are arranged, but we can't say what they are like in themselves.

The hiddenness of metaphysical simples, and our inability to describe what they are like, are, for Wittgenstein, not the result of remediable ignorance on our part. The mystery in which they are shrouded is essential to them, and closely connected with central doctrines of the *Tractatus*.

> 2.021 Objects form the substance of the world. Therefore they cannot be compound.

[20] Wittgenstein (1914–16), p. 18.

2.0231 The substance of the world *can* only determine a form and not any material properties. For these are first presented by the propositions—first formed by the configuration of the objects. [*The substance of the world* can *only determine a form, and not any material properties. For it is only by means of propositions that material properties are represented—only by the configuration of objects that they are produced.*]

2.0232 Roughly speaking: objects are colorless.

2.0233 Two objects of the same logical form are—apart from their external properties—only differentiated from one another in that they are different.

The first passage identifies objects with the substance of the world. The second tells us that this substance—the metaphysically simple objects—can only determine a form; they only have possibilities of entering into different configurations. In saying that they don't determine "material properties," Wittgenstein is, I take it, saying that they don't possess properties like shape or color; nor do the objects themselves determine which things have such properties. These properties are represented only by propositions; they come into being with "the configuration" of objects. In short, such properties are to be analyzed in terms of the relations among the simples.

In the third passage we are given an example. Colors are among the "material properties" that Wittgenstein is talking about. Since being a certain color—say red—is simply a matter of being made up of simples that stand in a certain configuration, the simples themselves aren't colored. Thus, we are told, they are colorless. Finally, in the fourth passage, two metaphysical simples of the same logical form—i.e., two simples with the same *possibilities* of combining with other objects—are said to have no *intrinsic* properties that differentiate them. They may have different external or *relational* properties; they may, as a matter of actual fact, happen to be combined with different objects, and so bear different relational properties. But apart from that there are no intrinsic properties to differentiate them. One of them, a, is simply different from, i.e., nonidentical with, b, whereas the other, b, is different from, i.e., nonidentical with, a.

Thus, for Wittgenstein the only thing we can say about simple objects is how they combine. He explicitly draws this conclusion at 3.221.

3.221 Objects I can only *name*. Signs represent them. I can only speak *of* them. I cannot *assert them*. A proposition can only say *how* a thing is, not *what it is*. [*Objects can only be* named. *Signs are their representatives. I can only speak* about *them: I cannot* put them into words. *Propositions can only say* how *things are*, *not* what *they are*.]

Although we can't say what metaphysical simples are like, we are supposed to be able to describe how they combine. But even this may be overoptimistic. Doctrines about necessity and possibility, which go to the

heart of the *Tractatus,* place severe constraints on the relational statements about metaphysically simple objects we can intelligibly make.

5. THE LOGICAL INDEPENDENCE OF ATOMIC SENTENCES AND ATOMIC FACTS

I have already highlighted the tractarian collapse of necessity and apriority into logical necessity. Various passages throughout the *Tractatus* contribute to this doctrine. For example, at 6.375 we are told that the only necessity is *logical* necessity and the only possibility is *logical* possibility.

> 6.375 As there is only a *logical* necessity, so there is only a *logical* possibility. [*Just as the only necessity that exists is* logical *necessity, so too the only impossibility that exists is* logical *impossibility.*]

From this we know that any proposition that is true at all possible world-states, and so is metaphysically necessary, is also a logical truth, and so is logically necessary. Since the converse is obvious, necessary truth and logical truth are the same. At 5.13, 5.131, and 4.1211 we are told that whenever propositions stand in any logical relation, they do so because of their *structure* (which is shown on an analysis that reveals their logical forms).

> 5.13 That the truth of one proposition follows from the truth of other propositions, we perceive from the structure of the propositions. [*When the truth of one proposition follows from the truth of others, we can see this from the structure of the propositions.*]
>
> 5.131 If the truth of one proposition follows from the truth of others, this expresses itself in relations in which the forms of these propositions stand to one another.
>
> 4.1211 If two propositions contradict one another, this is shown by their structure; similarly if one follows from another, etc.

This suggests the remarkable view that whenever q is a necessary consequence of p, a formal proof of q from p can be given; similarly, whenever p and q are necessarily inconsistent, the falsity of one can be formally derived from the truth of the other.

Two corollaries are (i) that one atomic proposition is never a necessary consequence of another—i.e., the truth of one atomic proposition never follows necessarily from the truth of another, and (ii) that atomic propositions are never incompatible with one another. Corollary (i) is made explicit in the sequence ending in 5.134.

> 5.132 If p follows from q, I can conclude from q to p; infer p from q. [*If p follows from q, I can make an inference from q to p, deduce p from q.*]
>
> The method of inference is to be made from the two propositions alone. [*The nature of the inference can be gathered only from the two propositions.*]

Only they themselves can justify the inference. [*They themselves are the only justifications of the inference.*]

Laws of inference, which—as in Frege and Russell—are to justify conclusions, are senseless and would be superfluous. [‘*Laws of inference’, which are supposed to justify inferences, as in the works of Frege and Russell, have no sense, and would be superfluous.*]

5.133 All inference takes place a priori. [*All deductions are made a priori.*]

5.134 From an elementary proposition no other can be inferred. [*One elementary proposition cannot be deduced from another.*]

In talking here about inference and deduction, Wittgenstein is talking about *a priori consequence*: q is an a priori consequence of p iff q can be validly deduced or inferred from p on the basis of a priori reasoning alone. Viewing such inference to be necessarily truth-preserving, he assimilated a priori consequence to necessary consequence and necessary consequence to logical consequence. Thus, we are told not only that no atomic proposition is a logical or a priori consequence of another, but also that no atomic proposition is a necessary consequence of another either. Corollary (ii) is explicitly endorsed at 6.3751 (c).[21]

6.3751(c) It is clear that the logical product of two elementary propositions can neither be a tautology nor a contradiction.

The idea behind these corollaries is clear. If an atomic sentence/proposition *Ha* logically entailed, or was logically incompatible with, another atomic sentence/proposition *Gb*, then the logical relation between the two would *not* be a matter of the structural relations between these two propositions, but rather would be about their subject matters, or contents. This cannot be so, because logic has no specific subject matter. Rather, the logical relationships holding among different sentences/propositions is always a purely formal matter; for Wittgenstein, it is always discoverable from an examination of their structure.

Since logic has no subject matter of its own, it has no method of finding out which atomic sentences/propositions are true and which are not. A central task of logic is to find sentences—logical truths, or tautologies—that are guaranteed to be true no matter how truth values are assigned to the atomic sentences; another task is to find sentences—contradictions—that are guaranteed to be false no matter how truth values are assigned to atomic sentences. Related to these tasks, logic will tell us when the truth of one sentence, or one set of sentences, guarantees the truth of another sentence, as well as when a set of sentences cannot jointly be true. If to

[21] Elementary propositions are atomic propositions. The logical product of two propositions is their conjunction. If the conjunction of two atomic propositions can never be a contradiction, then the two propositions cannot be incompatible.

this conception of logic one adds the tractarian doctrine that all necessity (and apriority) is logical necessity and all impossibility is logical impossibility, one gets the result that every necessary, and every a priori, truth is a logical truth, or tautology, and every necessary falsehood, and every proposition that can be known a priori to be false, is a logical falsehood or contradiction. One also gets the result that whenever the truth of one sentence/proposition necessitates the truth, or the falsity, of another, the second sentence/proposition is a logical consequence of the first, or *logically* incompatible with the first.

Suppose for the moment that Wittgenstein is right: if p and q are atomic sentences/propositions, then the truth, or the falsity, of p is always compatible with the truth, or the falsity, of q; it is possible for both to be true, both to be false, or either one to be true while the other is false. In short, the two are independent. Since the *Tractatus* posits a parallel between atomic sentences/propositions and atomic facts, the same sort of result holds for atomic facts. Thus, just after being told at 5.134 that one elementary proposition can never be logically deduced, or inferred, from another, we are given 5.135, while earlier we were given 2.061 and 2.062.

> 5.135 In no way can an inference be made from the existence of one state of affairs to the existence of another entirely different from it. [*There is no possible way of making an inference from the existence of one situation to the existence of another, entirely different from it.*]
>
> 2.061 Atomic facts are independent of one another.
>
> 2.062 From the existence or nonexistence of an atomic fact we cannot infer the existence or nonexistence of another.

These doctrines about the independence of atomic sentences/propositions and facts can be used to throw light on what atomic sentences/propositions really say about metaphysical simples, and what atomic facts really are possible. At 6.3751 (a,c), Wittgenstein provides an example of the kind of argument we can use.

> 6.3751(a) For two colors, e.g. to be at one place in the visual field, is impossible, logically impossible, for it is excluded by the logical structure of color.
>
> (c) (It is clear that the logical product of two elementary propositions can neither be a tautology nor a contradiction. The assertion that a point in the visual field has two different colors at the same time, is a contradiction.)

It follows from these remarks that there can be no meaningful atomic proposition *that a is red*, no proposition that says of some particular metaphysical simple that it is red. The reason that there can be no such atomic proposition is that if there were, its truth would be *incompatible* with the truth of the atomic proposition *that a is green*. Thus, the propositions—*that a is red* and *that a is green*—cannot be atomic. Likewise, there is no possible

atomic state of affairs that a is red, since this state of affairs would not be independent of the possible state of affairs that a is green.

This might not seem surprising, since we have already determined that, according to Wittgenstein, objects can't have color, or indeed any other material properties. But the point is much more far-reaching. Consider the following relational statements.

1a. a is to the right of b.
b. b is to the right of a.
c. a is to the right of a.
2a. a is heavier than b.
b. b is heavier than a.
c. a is heavier than a.
3a. a is exactly two inches away from b.
b. a is exactly one inch away from b.
c. a is exactly one inch away from a.
4a. a is touching b.
b. b is touching a.

In each case, the (a) and (b) statements are not independent of each other. In the first three cases they are incompatible with one another—i.e., it is impossible for both to be true. In the fourth case they are necessary consequences of one another—if one is true, then the other must be true. Similarly, statement (c) in the first three cases is necessarily false. These observations together with tractarian doctrines about atomic sentences/propositions entail that the statements in each example cannot all be atomic. Since in each example there is every reason to think that if one is atomic they all are, it follows from the fact that they are not logically independent that *none* qualify as atomic sentences, or propositions, in the sense of the *Tractatus*. We could produce the same sort of argument for virtually any statement involving spatial relations, temporal relations, relations involving measurement, or relations of relative size or degree. It follows that no statements of these types can be atomic sentences/propositions in the sense postulated by the *Tractatus*. This means that atomic sentences/propositions cannot attribute ordinary properties to metaphysical simples, nor can they attribute familiar relations involving space, time, measurement, or degree to these objects.

This leaves little or nothing we can imagine that atomic sentences/propositions can say. This is an incredible result. According to Wittgenstein, atomic sentences/propositions are the building blocks out of which all meaning is constructed. But if his doctrines are correct, we can scarcely conceive of any atomic sentences, or the specific contents they might have. In the end, he is forced into saying that all thought about the world reduces to thought about simple objects that have no properties we can identify and that can't be combined in any ways we can imagine; nevertheless they do combine in ways we can't comprehend. It is hard to understand what

this really amounts to, let alone why anyone should believe it. It is, I think, fair to say that few, if any, philosophers did.

Wittgenstein's views about metaphysical simples and the way they combine to form atomic facts are among the darkest and most implausible aspects of the *Tractatus*. But other aspects of the *Tractatus* were much more interesting and influential. Particularly important were the doctrines about the nature of truth, meaning, and propositions, as well as related doctrines about logic, necessity, possibility, and the relationship between logically complex and atomic sentences/propositions. These aspects of the *Tractatus* are examined in the next two chapters.

CHAPTER 2

The Single Great Problem of the *Tractatus*: Propositions

1. PICTURES, REPRESENTATIONS, AND LOGICAL FORM

In the previous chapter I discussed Wittgenstein's conception of metaphysical simples and the way they combine to form atomic facts. I now turn to his views on truth, meaning, propositions, necessity, possibility, conceivability, and logic. As before, I begin with atomic sentences, which, we are told, are combinations of logically proper names that *picture* or represent possible states of affairs.[1]

2.01 An atomic fact is a combination of objects.

2.1 We make to ourselves pictures of facts. [*We picture facts to ourselves.*]

2.11 The picture presents the facts in logical space, the existence or non-existence of facts. [*A picture presents a situation in logical space, the existence and non-existence of states of affairs.*]

[1] In citing passages from the *Tractatus*, I use the Ogden translation. When useful I add the Pears and McGuinness translation (Wittgenstein 1922 [1961]), italicized, in square brackets.

2.12 A picture is a model of reality.
3.1 In the proposition the thought is expressed perceptibly through the senses.
3.12 The sign through which we express the thought I call the propositional sign. And the proposition is the propositional sign in its projective relation to the world.
3.2 In propositions thoughts can be so expressed that to the objects of the thoughts correspond the elements of the propositional sign.
3.201 These elements I call "simple signs" and the proposition "completely analyzed".
3.202 The simple signs employed in propositions are called names.
4.0311 One name stands for one thing, another for another thing, and they are connected together. And so the whole, like a living picture, presents the atomic fact.

In the tractarian system, each name designates a single object, which is its meaning, and each object is named by a single name. The way names are arranged in an atomic sentence represents a way in which the objects they name could be combined. Atomic sentences, which Wittgenstein sometimes calls *elementary propositions*, are said to picture possible facts (or states of affairs) that might (informally) be taken to be their meanings.

2.13 To the objects correspond in the picture the elements of the picture. [*In a picture objects have the elements of reality corresponding to them.*]
2.131 The elements in the picture stand, in the picture, for the objects. [*In a picture the elements of the picture are the representatives of objects.*]
2.201 The picture depicts reality by representing a possibility of existence and non-existence of atomic facts.
2.202 The picture represents a possible state of affairs in logical space.
2.221 What a picture represents is its sense.
3.203 A name means an object. The object is its meaning.
3.22 In the proposition, the name represents the object.

The picture analogy is central to Wittgenstein's conception of meaning. The analogy can be illustrated by a pair of well-known examples. The first is a courtroom model of a traffic accident in which toy cars stand for real cars. In the model, putting the toy cars in a certain spatial arrangement represents real cars as being in that arrangement. In this example the spatial properties and relations of the model allow it to picture or represent spatial properties or relations of the real cars. The second example is a painting of a barn. By making a certain portion of the canvas red, one represents the barn as red.

This is what Wittgenstein has in mind at 2.171.

2.171 The picture can represent every reality whose form it has. The spatial picture, everything spatial, the colored, everything colored, etc. [*A*

picture can depict any reality whose form it has. A spatial picture can depict anything spatial, a colored one anything colored, etc.]

In language, we don't use different colored inks to represent the different colors of the referents of our words. Nor do we place words in spatial relationships that directly correspond to the spatial relationships existing among the items we are talking about. Still, Wittgenstein thought, an atomic sentence represents a nonlinguistic fact, or state of affairs, only by sharing a common form with it. This common form can't always be a spatial one, as in the traffic model, or a material one involving properties like color, as in the case of the painting of the barn. Thus, he says that the form shared by an atomic sentence and the state of affairs it represents must be an abstract *logical form*.[2]

2.161 In the picture and the pictured there must be something identical in order that the one can be a picture of the other at all. [*There must be something identical in a picture and what it depicts to enable the one to be a picture of the other at all.*]

2.17 What a picture must have in common with reality, in order to be able to represent it after its manner—rightly or falsely—is its form of representation. [*What a picture must have in common with reality, in order to be able to depict it—correctly or incorrectly—in the way that it does, is its pictorial form.*]

2.18 What every picture, of whatever form, must have in common with reality, in order to be able to represent it at all—rightly or falsely—is the logical form, that is, the form of reality.

2.181 If the form of representation is the logical form, then the picture is called a logical picture. [*A picture whose pictorial form is logical form is called a logical picture.*]

2.182 Every picture is *also* a logical picture. (On the other hand, not every picture is, for example, a spatial one.)

3. The logical picture of the facts is the thought.

This doctrine is not as mysterious as it sounds. For Wittgenstein, an atomic sentence is a *linguistic fact,* a structured combination of names, while a state of affairs is a *nonlinguistic fact*—a structured combination of objects. In order for the linguistic fact to represent nonlinguistic reality, something about the way the names are combined in the sentence must correspond to the way the objects are combined in the state of affairs.

3.14 The propositional sign consists in the fact that its elements, the words, are combined in it in a definite way.
The propositional sign is a fact.

3.142 Only facts can express a sense, a class of names cannot.

[2] See chapter 2 of Fogelin (1987) for further discussion of the picture theory.

3.1432 We must not say, "The complex sign '*aRb*' says '*a* stands in relation *R* to *b*'"; but we must say, "*That* '*a*' stands in a certain relation to '*b*' says *that aRb*".

3.2 In propositions thoughts can be so expressed that to the objects of the thoughts correspond the elements of the propositional sign.

3.21 The configuration of the simple signs in the propositional sign corresponds to the configuration of objects in the state of affairs.

There is not much to be said, in a general way, about the required correspondence between the relation in which the names stand in the sentence and the relation in which the objects named are represented as standing in the world. The correspondence is mostly a matter of linguistic convention.

It is a convention of language users that certain ways of combining names—i.e., certain ways of placing them in relations to one another to form a sentence—represent the named objects as standing in certain relations. For example, in

1. Los Angeles is south of San Francisco.

the name 'Los Angeles' stands in the relation ___*immediately preceding the words 'is south of', which immediately precede*___ to the name 'San Francisco'. Placing the names in this syntactic relation represents the object Los Angeles as standing in the relation *being to the south of* to the object San Francisco. Speakers of a language adhere to linguistic conventions specifying (i) which objects different names designate, and (ii) which nonlinguistic relations (holding between the objects) the linguistic relations (holding between the names in the sentence) stand for. When Wittgenstein says that an atomic sentence and an atomic fact share a logical form, he means (i) that just as the atomic fact is a complex in which objects stand in a relation R_o, so the atomic sentence is a fact in which names stand in a relation R_n, and (ii) that linguistic conventions stipulate which objects are designated by which names, and which relation R_o the objects are represented as standing in by the use of R_n to relate the names.

To recap, in order for one fact to stand in a representational relationship to another fact, the two facts must share a common form. Sometimes, as with the traffic model and the representational painting, that form involves material properties and relations, colors or spatial relations, being common to the two facts. In other cases, as with language, the common form is simply an abstract logical form involving a conventional correlation.

2. TRUTH, MEANING, AND TRUTH FUNCTIONALITY

According to the picture theory, what makes atomic sentences representational, and hence meaningful, is similar to what makes some paintings representational. What makes the painting representational is not the

existence of any actual thing that is represented; a painting of a winged horse can be representational, even though there are no winged horses. What makes a painting representational is, plausibly, that it depicts something that could, or might conceivably, exist, even if in fact it doesn't. Similarly, an atomic sentence is representational, and hence meaningful, if and only if it is possible for the objects designated by its names to be related as they are represented to be by the way the names in the sentence are related. If one thinks that, in addition to actual facts, some facts are merely possible, one could say that the meaning of an atomic sentence is the *possible* fact it represents as existing.

That wasn't Wittgenstein's view. For him, nothing is both a fact and merely possible. Indeed, nothing is the meaning of an atomic sentence. Such sentences are meaningful, but no entities are their meanings. An atomic sentence S is meaningful if and only if the objects designated by its names *could* stand in the relation conventionally indicated by the relation in which they stand in S. To understand S is *not* to be acquainted with an abstract entity—a meaning, a Fregean thought, a Russellian proposition, or a possible state of affairs. It is to know what the world would have to be like if S were to be true.

> 4.024 To understand a proposition means to know what is the case if it is true.
> (One can therefore understand it without knowing whether it is true or not.)

Here and throughout Wittgenstein's uses of 'proposition' approximate his uses of 'meaningful sentence'.

This picture comes out in the following passages.

> 3.11 We use the sensibly perceptible sign (sound or written sign, etc.) of the proposition as a projection of the possible state of affairs. The method of projection is the thinking of the sense of the proposition.
>
> 3.12 The sign through which we express the thought I call the propositional sign. And the proposition is the propositional sign in its projective relation to the world.
>
> 3.13 To the proposition belongs everything which belongs to the projection; but not what is projected. [*A proposition includes all that the projection includes, but not what is projected.*]
> Therefore the possibility of what is projected but not this itself. [*Therefore, though what is projected is not itself included, its possibility is.]*
> In the proposition, therefore, its sense is not yet contained, but the possibility of expressing it. [*A proposition, therefore, does not actually contain its sense, but does contain the possibility of expressing it.*]

A proposition is a "propositional sign," i.e., a sentence, "*in its projective relation to the world.*" It is a sentence *used* "as a projection of a possible situation." Being a meaningful sentence, it has a sense. At 2.202 and 2.221

we were told that a picture represents a possible fact (state of affairs) that is its sense. Now we are told that "what is projected," the possible fact, is *not* included in the proposition, but "its possibility is." It can't be included, because there are no facts for false propositions to contain, and because we must grasp the sense of a proposition before we know whether it is true.

Even though the possible fact isn't "included" in the proposition, "the method of projection"—"*the thinking of the sense of the proposition*"—is. What is this method? To think of the sense of the proposition is to use the propositional sign in accord with the conventions governing its names and the syntactic structure in which they stand. These conventions determine what fact must exist if the sentence, so used, is to be true. The language is obscure, but the idea isn't; linguistic conventions are *somehow* included in the proposition, qua meaningful sentence, as what one must know to understand it.

What about non-atomic propositions? For Wittgenstein, the truth or falsity of any non-tautological, noncontradictory proposition is determined by its relation to the world, which is the totality of atomic facts.

1. The world is everything that is the case.
1.1 The world is the totality of facts, not of things.
1.12 For the totality of facts determines both what is the case, and also what is not the case.
2.04 The totality of existent atomic facts is the world.
2.05 The totality of existent atomic facts also determines which atomic facts do not exist.

Thus the truth or falsity of all meaningful, non-tautological, noncontradictory, propositions is determined by the totality of atomic facts.

4.2 The sense of a proposition is its agreement and disagreement with the possibilities of the existence and non-existence of the atomic facts.
4.21 The simplest proposition, the elementary proposition, asserts the existence of an atomic fact.
4.25 If the elementary proposition is true, the atomic fact exists; if it is false, the atomic fact does not exist. [*If an elementary proposition is true, the state of affairs exists: if an elementary proposition is false, the state of affairs does not exist.*]
4.26 The specification of all true elementary propositions describes the world completely. The world is completely described by the specification of all elementary propositions plus the specification of which of them are true and which false. [*If all true elementary propositions are given, the result is a complete description of the world. The world is completely described by giving all elementary propositions, and adding which of them are true and which false.*]

It follows that there are no negative, disjunctive, or other non-atomic facts to which true, truth-functionally-complex propositions correspond. There

are no complex facts to which complex sentences of any sort correspond. Rather, the truth or falsity of a non-atomic sentence is *always* determined by the truth or falsity of atomic propositions. So, the negation of a false atomic proposition is true, not because it corresponds to a negative fact, but because the true proposition it negates corresponds to no fact.

Wittgenstein adopts a two-stage theory of meaning and truth. At stage 1, atomic sentences are declared meaningful if and only if the relations in which they represent objects as standing are relations in which the objects could stand; they are true if and only if there are facts in which the objects do stand in those relations. At stage 2, non-atomic sentences are declared meaningful and true on the basis of the truth and meaning of related atomic sentences. There are no conventions specifying the properties that uses of non-atomic sentences represent objects as having, and no non-atomic facts to which non-atomic truths correspond. For example, to know the meaning of a negative sentence ⌜~S⌝ is to know the meaning of S, and to understand the negation operator. That operator doesn't name anything, and ⌜~S⌝ doesn't picture a negative fact. [3]

> 4.0312 The possibility of propositions is based upon the principle of the representation of objects by signs. [*The possibility of propositions is based on the principle that objects have signs as their representatives.*]
> My fundamental thought is that the "logical constants" do not represent. That the *logic* of facts cannot be represented. [*My fundamental idea is that the 'logical constants' are not representatives; there can be no representatives of the logic of facts.*]
>
> 5.44(e) And if there was an object called '~', then '~ ~ *p*' would have to say something other than '*p*'. For the one proposition would then treat of ~, the other would not. [*And if there were an object called '~', then it would follow that '~ ~ p' said something different from what 'p' said, just because the one proposition would then be about ~ and the other would not.*]

To know the meaning of ⌜~S⌝ is to correlate it with S, which represents certain objects as being related in certain ways. So if S represents $o_1 \ldots o_n$ as standing in relation R, then to understand ⌜~S⌝ is to know that it is true if and only if S isn't true, if and only if $o_1 \ldots o_n$ don't stand in relation R. Similar remarks hold for other truth-functionally compound sentences.

For Wittgenstein, atomic facts are all the facts there are. If there were possible facts, we would say that different combinations of possible atomic facts constitute different possible world-states. A better way of putting this is to say that there is nothing to any such state over and above the atomic facts that would exist if the world were in that state. We can also express this linguistically. Let A be the set of all atomic sentences, and f be an assignment of truth values to its members. The set of sentences to

[3] See pp. 69–72 and 177–84 of Black (1964).

which f assigns truth represents one possible world-state. If we had a different assignment, f', the set of sentences to which f' assigned truth would represent a different world-state. Finally, consider every possible assignment of truth values to members of A—i.e., every distribution of truth and falsehood to atomic sentences. One assigns truth to every atomic sentence, one assigns falsity to every atomic sentence, and for every possible combination between these two extremes, there will be an assignment that gives that combination of truth values to the sentences in A. The doctrines of the *Tractatus* maintain that each assignment represents a possible world-state (a way the world could have been), and that each possible world-state is represented by an assignment.

> 4.27 With regard to the existence of n atomic facts there are 2^n possibilities. It is possible for all combinations of atomic facts to exist, and the others not to exist.
> 4.28 To these combinations correspond the same number of possibilities of the truth—and falsehood—of n elementary propositions.
> 4.3 The truth possibilities of the elementary propositions means the possibilities of the existence and non-existence of the atomic facts.
> 4.4 A proposition is an expression of agreement and disagreement with the truth-possibilities of the elementary propositions.
> 4.41 The truth possibilities of the elementary propositions are the conditions of the truth and falsehood of the propositions.
> 5(a) Propositions are truth functions of elementary propositions.

Suppose S is non-atomic. Since we know that there are no non-atomic facts, we know that S's truth value can't consist in its correspondence, or lack of correspondence, with a non-atomic fact. Rather, S's truth value must be determined by which atomic facts exist, or equivalently, by which atomic sentences are true and which are false. According to the *Tractatus, every proposition (meaningful sentence) is a truth function of atomic propositions*. So, any assignment of truth values to all meaningful atomic sentences (propositions) determines the truth value of every proposition. Wittgenstein also thought that to know the meaning of any logically complex sentence is to know how its truth or falsity is determined from atomic sentences.

Although this approach is attractive, it imposes restrictions on the relationship between truth and meaning that Wittgenstein didn't clearly recognize. If understanding S is knowing its truth conditions, then understanding the truth predicate can't depend on antecedently understanding S. Moreover, the claims made by S and ⌜'S' is true⌝ can't be a priori consequences of one another. If they were, we would get the absurd result that knowing *that snow is white if and only if snow is white* is sufficient for knowing that *'snow is white' is true if and only if snow is white*, and hence for knowing the truth conditions of the sentence 'snow is white'. But surely, knowing that *snow is white if and only if snow is white* tells us nothing about the

meaning of 'snow is white'. Finally, the truth-functional connectives can't all be defined in terms of truth and falsity, since any such definition—e.g., ⌜~S⌝ is true if and only if S is not true—will presuppose one or more of the connectives to be defined. These limitations will become important when we look more closely at the details of Wittgenstein's account of truth and propositions.

3. MEANINGFULNESS WITHOUT MEANINGS

Wittgenstein's theory of meaning in the *Tractatus* was rightly seen as a new departure. Like Frege and Russell, Wittgenstein took sentences to be the primary meaning-bearing units, while taking the significance of subsentential expressions to consist in the contributions they make to the meanings of sentences in which they occur. But unlike Frege, the early Moore, and the early Russell, he did not identify what it is for a sentence to be meaningful with its expressing, or standing for, a nonlinguistic entity that is its meaning. For him there is nothing that is the meaning of a meaningful sentence. This idea, introduced in the *Tractatus*, was to have long and lasting influence, some which continues to this day.[4]

The view previously endorsed by Frege, Moore, and Russell, but repudiated in the *Tractatus,* was that ordinary declarative sentences express propositions that are (i) the meanings (semantic contents) of those sentences (in contexts of use), (ii) the referents of embedded clauses ⌜that S⌝, (iii) the primary bearers of truth and falsity, and hence that in virtue of which (uses of) sentences are true or false, and (iv) the things an agent A asserts, believes, or knows when ⌜x asserts/believes/knows that S⌝ is true of A. Although there is much to be said in favor of this view, there was enough to be said against their versions of it that Moore and Russell abandoned propositions between 1910 and 1912.[5] As indicated in volume 1, Russell tried to dispense with propositions in favor of his "multiple relation theory judgment," but the difficulties with this theory were overwhelming, and by 1919 he had abandoned it.[6]

Wittgenstein's new approach took (something like) *meaningful sentences* to be the primary bearers of truth and falsity, while insisting that for a sentence to be meaningful was *not* for there to be anything that was its meaning. Unlike Frege, Russell, and Moore, Wittgenstein didn't specify what sentential clauses ⌜that S⌝ refer to, or what one asserts, believes, or knows

[4] Although a similar idea is expressed in Russell (1918–19), Russell was then following the then unpublished Wittgenstein (1914–16). For references, see the notes on pp. 571–76 of Soames (2014), volume 1 of this work.

[5] See (Soames 2014): chapter 2, section 2; chapter 3, section 3; chapter 7, sections 4.1, 4.2, 4.5; chapter 8, section 2.3.5; chapter 9, sections 2–6; chapter 12, section 2.

[6] See Soames (2014), chapters 9, 10, and section 6 of chapter 12.

when one asserts, believes, or knows something. Because the earlier theories had answered these questions, they were able to identify what truth is predicated of in examples like those in (2), (3), and (4).

2a. It is true *that the earth is round*.
b. *That 2 + 2 = 4* is necessarily true.
c. The proposition *that 'Wittgenstein' names Wittgenstein* is only contingently true.
3a. Every proposition is such that it or its negation is true.
b. Some propositions advanced in the *Tractatus* are true.
4a. What Mary asserted is true.
b. The proposition Mary asserted is true, despite the fact that Bill denied it.

Because Wittgenstein didn't address many questions that Frege, Russell, and Moore used nonlinguistic propositions to answer, it's not obvious what should be said on his behalf about these examples. Still, it is reasonably clear where to begin. Since he seemed to identify propositions with (something like) meaningful sentences, which he characterized as the bearers of truth value, one would expect him to take whatever corresponds to the *that*-clauses in (2) to designate the sentences used there, while also expecting meaningful sentences to be the targets of predication in examples like (3). The sentences in (4) raise further complications that will be explored later.

Suppose meaningful sentences are bearers of truth conditions. For them to be *meaningful* is for them to be used in a certain way. But what, exactly, are sentences *used in a certain way*? Let 'Los Angeles' and 'San Francisco' be names and sentence (1)

1. Los Angeles is south of San Francisco.

be the tractarian linguistic fact that consists of the former name standing in the two-place relation R—*immediately preceding 'is south of', which is followed by*—to the latter name. What are its truth conditions? The answer depends on what the sentence means, which in turn depends on conventions governing its use. Let the conventions governing the names be that 'Los Angeles' is to be used as a logically proper name for the city Los Angeles and that 'San Francisco' is to be used as a logically proper name for the city San Francisco. Let the convention governing the relation R be that structures in which two names stand in this relation are *to represent* the object designated by the first name as *being located to the south of* the object designated by the second. Given these conventions, one who *uses* (1) predicates *being to the south of San Francisco* of Los Angeles, thereby representing the latter as being south of the former.

It is because of these conventions that the sentence is true iff Los Angeles is south of San Francisco. But what exactly is this truth bearer? Following Wittgenstein, I have taken the sentence to be the tractarian propositional sign—which is the linguistic fact in which the name 'Los Angeles' bears the syntactic relation R to the name 'San Francisco'. This specification of

the fact with which sentence (1) is identified doesn't mention the conventions governing either the names or R. Thus, it is natural to suppose that although the sentence *is* governed by these conventions, it didn't have to be. Had it been governed by other conventions, it would have meant something other than what it does mean, and so had different truth conditions from those it actually has.

Viewed in one way, this is no surprise. There is something—the bare syntactic structure I have sketched—that *is* used by speakers to represent Los Angeles as being south of San Francisco, but could have been used differently, and so had different truth conditions. Although it has these truth conditions contingently, surely there is something else that has them essentially. We do say that *necessarily, the* proposition *that Los Angeles is south of San Francisco is true iff Los Angeles is south of San Francisco*. Since we couldn't affirm this if the *proposition* were a tractarian *propositional sign*, it would seem that the proposition isn't the propositional sign.

Let's make sure by supposing that it is and deriving a falsehood. Suppose further that the sentence/propositional sign could have been governed by a convention other than the one that actually governs it and that a sentence that represents objects as standing in a certain relation (while representing nothing else) is true if and only if the objects do stand in that relation. We then appeal to the following.

MODAL SCHEMATA

It could have been the case that such-and-such was (were) so-and-so if and only if possibly such-and-such is (are) so-and-so, if and only if for some possible world-state w, such-and-such is (are) so-and-so at w.

It couldn't have been the case that such-and-such wasn't (weren't) so-and-so if and only if necessarily such-and-such is (are) so-and-so, if and only if for all possible world-states w, such-and-such is (are) so-and-so at w.

Such-and-such is (are) so-and-so at w if and only if were w actual, such-and-such would be so-and-so.

In the presence of these schemata, our assumptions (including the assumption that sentences, or propositional signs, are propositions) are inconsistent with the obvious truths (5a) and (5b).

5a. Necessarily, the proposition that Los Angeles is south of San Francisco is true if and only if Los Angeles is south of San Francisco.

b. For all world-states w, the proposition that Los Angeles is south of San Francisco is true at w if and only if at w, Los Angeles is south of San Francisco.

Consider a world-state w*, geographically similar to the actual world-state, at which sentence (1) (the bare syntactic structure) represents San Diego as being east of San Antonio, even though, at w*, San Diego is west of San Antonio. Appealing to the third modal schema, with sentence (1) playing the role of 'such-and-such' and 'true' playing the role of 'so-and-so', we derive R.

R. If w* were actual, then sentence (1) would be *false*, even though Los Angeles was south of San Francisco.

This can't be, if (5a) and (5b) are true and the proposition that Los Angeles is south of San Francisco is identical with the sentence, as we have conceived it.

Some may be tempted to try to avoid this result by modifying the third modal schema when predications of truth or falsity are involved. The idea is to *define* what it is for a sentence S to be true at w as follows.

A NEW TWO-PLACE TRUTH PREDICATE

For all world-states w and sentences S, S is true at w if and only if (i) S actually represents things to be a certain way (i.e., S does so at the actual world-state), and (ii) if w were actual, things would be that way.

There are three reasons to resist this. First, there is no rationale (apart a desire to save the dubious thesis that propositions are syntactically individuated sentences) for adopting a special account of modal predications of *truth* that differs from the accepted account of modal predications of other properties.[7] Second, to do so would be to obliterate an obviously correct result; *if* the sentence 'Los Angeles is south of San Francisco' *were* governed by the possible conventions associated with w* above, then it *would* be false at w* because it *would represent* San Diego as being east of San Antonio. The third problem arises from asking what any possible agent must know in order to know that John *asserted* the proposition that Los Angeles is south of San Francisco. If the proposition were identical with the sentence 'Los Angeles is south of San Francisco', then what *any possible agent* would have to know is that John assertively uttered some sentence that represents what the sentence 'Los Angeles is south of San Francisco' *represents at* @—the state the world is actually in. But *merely possible* agents don't have to know anything about @ in order to know that John asserted that Los Angeles is south of San Francisco.[8]

For all these reasons one should be wary of taking propositions to be tractarian propositional signs. Is there a more adequate view that preserves important tractarian themes? The natural alternative is to take tractarian propositions to somehow combine bare propositional signs with the conventions that govern their use. With this in mind, let the propositional sign be (1) and the conventions governing its use be (a) that 'Los Angeles' is to be used as a logically proper name for the city Los Angeles, (b) that 'San Francisco' is to be used as a logically proper name for the city San Francisco, and (c) that structures consisting of one name standing in the relation ___*immediately precedes 'is south of', which itself immediately precedes*___ to

[7] See Soames (2010b).

[8] The logic of arguments of this form is given in chapter 2 of Soames (2002).

another name are to be used *to represent* the object designated by the first name as *being to the south of* the object designated by the second name. We may then identify the proposition that Los Angeles is south of San Francisco with *a use* of the structured syntactic form (1) in accord with these conventions *to represent Los Angeles as being south of San Francisco*.

In order for this to work, there must be an entity of some kind—*a use of the sentence in accord with these conventions*—of which we predicate truth. What is such a use? Since to use the sentence is to do something, *a use of the sentence* is a type of cognitive act—one performed by different agents who use the sentence in the same way. In our example, it is *the act of using the names, 'Los Angeles' and 'San Francisco', to designate the cities, Los Angeles and San Francisco, while using the relation the names stand in to represent the item designated by the first name as being south of the item designated by the second*. The use—*the act (type) itself*—represents Los Angeles as being south of San Francisco, in the sense that *any possible agent* who performs it thereby represents Los Angeles to be south of San Francisco. Since to do that is to represent the two cities accurately, the use may naturally be said to be true. Moreover, it has its truth conditions essentially.[9]

This reconstruction preserves central tractarian themes. (i) It explains the meaningfulness of the sentence without positing an independent entity as its meaning. (ii) It identifies the truth-bearer, *the meaningful use*, as an entity the truth of which is defined in terms of its representational accuracy. (iii) It stipulates that the constituents of the sentence, the names and the syntactic relation, are isomorphic to the constituents of the atomic fact that makes a use of it true. (iv) It recognizes that the conventions governing the use of the sentence are those governing its sub-sentential constituents; no extra convention governing the sentence as a whole is needed. (v) It maintains that the proposition has its truth conditions essentially because any *possible* agent using the sentence in this way represents Los Angeles as being south of San Francisco.

All this is as it should be, but it isn't exactly what Wittgenstein had in mind. Although uses of sentences represent, or picture, reality, his propositional pictures are *facts*—not *acts*.

> 2.14 The picture consists in the fact that its elements are combined with one another in a definite way.
> 2.141 The picture is a fact.
> 2.21 The picture agrees with reality or not; it is right or wrong, true or false.
> 2.221 What a picture represents is its sense.
> 2.222 In the agreement or disagreement of its sense with reality, its truth or falsity consists. [*The agreement or disagreement of its sense with reality constitutes its truth or falsity.*]

[9] See chapter 2 of Soames (2015b).

What facts does he have in mind? His propositional signs are facts. Could *they* be tractarian propositions? He does say that propositions are perceptible.

3.1 In the proposition the thought is expressed perceptibly through the senses.

But he also says things that distinguish propositions from propositional signs.

3.11 We use the sensibly perceptible sign (sound or written sign, etc.) of the proposition as a projection of the possible state of affairs. [*We use the perceptible sign of a proposition (spoken or written, etc.) as a projection of a possible situation.*] The method of projection is the thinking of the sense of the proposition.

3.12 The sign through which we express the thought I call the propositional sign. And the proposition is the propositional sign in its projective relation to the world.

3.13 To the proposition belongs everything which belongs to the projection; but not what is projected. [*A proposition includes all that the projection includes, but not what is projected.*]

Therefore the possibility of what is projected but not this itself. [*Therefore, though what is projected is not itself included, its possibility is.*]

In the proposition, therefore, its sense is not yet contained, but the possibility of expressing it. [*A proposition, therefore, does not actually contain its sense, but does contain the possibility of expressing it.*]

("The content of the proposition" means the content of the significant proposition.) [*'The content of a proposition' is the content of a proposition that has sense.*]

In the proposition the form of its sense is contained, but not its content. [*A proposition contains the form, but not the content, of its sense.*]

3.14 The propositional sign consists in the fact that its elements, the words, are combined in it in a definite way.

The propositional sign is a fact.

It appears from these passages that tractarian propositions are not identical with propositional signs. The latter are bare syntactic structures which, though they may be meaningful, aren't individuated by what they mean. It is tempting to say, as some passages seem to suggest, that the *sense of the proposition* is the *possible fact* that consists of the objects designated by its names being related as they are represented to be. In other words, the sense of the proposition is the nonlinguistic fact that would make the proposition true, were that fact actual. But this wasn't Wittgenstein's view. For him nothing is both a fact and *merely possible*. He registers this obliquely by saying that propositions don't *contain* their senses. They can't because there are no facts for false propositions to contain, and because we must grasp the sense of a proposition before we know whether it is true.

Recall his words. "The method of projection is the thinking of the sense of the proposition." In thought, the proposition we entertain represents worldly items—the objects that are projections of the names in the propositional sign—as standing in the relation that is the projection of the relation R that unites the names in the propositional sign. We are told that the proposition "includes all that the projection includes, but not what is projected." This last item, *what is projected,* is the sense of the proposition—the possible fact. It isn't "included" in the proposition, nor are the objects and relations that are projections of the constituents of the propositional sign. But the rest of the projection is included. What are the remaining items? They must be whatever elements are responsible for determining what the names and the syntactic relation R project; they are the conventions governing the names plus the convention governing R. They are needed to determine what fact would have to exist if the proposition were true. These conventions, which *aren't* included in the propositional sign, *are* somehow included in the proposition as what one must know in order to understand its representational content.

How are they included? The propositional sign is a purely syntactic structure in which symbols stand in a certain relation. Wittgenstein tries to identify the proposition using the phrase *the propositional sign in its projective relation to the world.* Unfortunately, this language, *the sentence S in its relation to the world,* doesn't pick out an entity other than S—any more than the phrases *Scott-in-his-relation-to-this-book, Scott-in-his-relation-to-USC, Scott-in-his-relation-to-his-wife,* and so on pick out entities other than me of which I am, nevertheless, an essential part. There aren't several Scotts, or Scott-complexes, here, just misleading ways of talking about the fact that I am the author of this book, I teach at USC, and I live with my wife. The same is true of Wittgenstein's talk of propositional signs *in their projective relations to the world*.

Wittgenstein's confused terminology parallels the familiar contemporary terminology contrasting *interpreted* and *uninterpreted* sentences. As applied to a language that is actually used, these terms don't designate two kinds of sentences; they signify two ways of speaking about the same sentences.[10]

[10] Black (1964) perpetuates Wittgenstein's error on p. 98, where tractarian propositions, thought of as meaningful sentences, are contrasted with "uninterpreted sentences." This is repeated on p. 99, where the following passage occurs: "The word *Satz* is used in German to stand for what we should call a 'sentence' as well as for what we would call a 'proposition' (or 'statement' . . .). Wittgenstein sometimes distinguishes the two senses by using 'propositional sign' (*Satzzeichen*, 3.12a) for the sentence. . . . It is essential to Wittgenstein's conception that the proposition should be expressed *in* a sentence. . . . A disembodied proposition would be an absurdity. Thus it is natural for him to use *Satz* to cover both aspects—the perceptual sign and its sense. . . . [I]t is essential to a proposition that it makes an abstract truth-claim." Essentially the same confusion occurs in Black's discussion on pp. 81–82 of a "picture-vehicle" and "a picture in the full sense when its elements have been co-ordinated in

An "uninterpreted sentence" is a syntactic structure, a kind of linguistic fact. An "interpreted sentence" is a meaningful use of a sentence, a kind of cognitive act. Wittgenstein rightly denied that propositions are propositional signs, but he failed to identify them with any genuine entities, while making it seem as if his pseudo-entities were the only candidates. There are, of course, *artificial* ways of combining conventions governing meaningful uses of sentences with the sentences they govern into a single entity that can go proxy for genuine propositions. For some purposes, ordered pairs of conventions and propositional signs will do. But they aren't propositions. Propositions have their representational properties and truth conditions inherently; the pairs don't, but rather require interpretation by us. The solution to Wittgenstein's problem is to take propositions to be *uses of sentences* in accord with conventions.

Had he done so, he would also have had to rethink his use of the truth predicate so as to recognize the a priori equivalence of the claims (6a) and (6b) and the lack of such equivalence between the claims (7a) and (7b).

6a. The proposition that Los Angeles is south of San Francisco is true.
b. Los Angeles is south of San Francisco.
7a. The sentence 'Los Angeles is south of San Francisco' is true.
b. Los Angeles is south of San Francisco.

As noted in section 2, Wittgenstein needs a conception of truth applying to meaningful sentences according to which knowledge of their truth conditions is knowledge of their meanings. For this S and ⌜'S' is true⌝ cannot be a priori equivalent. But as I have here indicated, he also needs a notion of truth according to which propositions have their truth conditions essentially. Unfortunately, as Black illustrates, he muddles these together by taking the truth bearers to be sentences while assuming an a priori equivalence that requires non-sentential bearers.

> According to Wittgenstein's conception, the proposition expressed by the sign '*p* is true' has exactly the same truth conditions as the proposition expressed by '*p*', and is therefore identically the same proposition (cf. 5.141). There is no way of interpreting 'p is true' as a truth function of 'p' that does not identify it with 'p'. As he says in the *Notebooks* (9 (7)), '"p" is true' says (*aussagt*) nothing else but 'p'. . . . Similarly, '"*p*" is false' says the same, is exactly the same proposition as, 'not-*p*'.[11]

The tractarian confusion about truth will be important when we look more carefully at truth-functionally complex propositions.

a determinate way with objects, upon the understanding that those objects are supposed to be connected as their proxies are in fact connected in the vehicle."
[11] Black (1964), p. 218.

4. PROPOSITIONS

4.1. The First Misstep: Symbolic Artifacts vs. What We Do with Them

The first crucial misstep in Wittgenstein's solution to "the single great problem of philosophy" was his (qualified) identification of propositions with symbolic artifacts of representational systems. This led him to take propositions—thought of as the fundamental units of representation and primary bearers of truth and falsity—to be sentences, rather than uses of sentences. There are two main problems with this idea. First, propositions can't be identified with bare syntactic forms (tractarian propositional signs); nor can they be composite entities consisting of such forms plus the conventions governing them. Second, there is nothing of significance that is *essential* to all and only those artifacts that can be used to represent reality other than the fact that they are, or can be, so used. What is essential to thought is that *agents* represent things as being certain ways, not what, if any, instruments they use in doing so.

Any organism whose cognitions can be true or false represents things as being various ways. Sometimes it does so by using symbols. Thus, *some propositions* may be uses of sentences to represent things as being certain ways. But there is no need to suppose that an agent *always* uses symbols when thinking of something as dangerous, or perceiving one thing to be bigger than something else. Agents perform many kinds of representational cognitive acts. Sometimes they do so linguistically, in which case (some) propositions they affirm may be uses of symbols. Sometimes they represent things as being certain ways nonlinguistically, in which case the propositions they affirm or believe aren't uses of symbols.

Thus, I reject what Wittgenstein says at 4.0312.

> 4.0312 The possibility of propositions is based upon the principle of the representation of objects by signs. [*The possibility of propositions is based on the principle that objects have signs as their representatives.*]

About this Max Black says:

> It is essential to Wittgenstein's conception that the proposition should be expressed in a sentence. . . . A disembodied proposition would be an absurdity.[12]

But why should we think that representation is inherently linguistic? If one kind of cognitive act—a use of a sentence—can represent things accurately or inaccurately, and so be true or false, why can't the same be said of related cognitive acts, in which we nonlinguistically *perceive*, *imagine,* or *think of* things as being certain ways? If I am right, the tractarian insistence on symbolic representation misrepresents the essence of representational thought.

[12] Ibid., p. 99.

In one way, this criticism takes us a step toward Wittgenstein's later philosophy, where he rejects the idea that the essence of thought lies in the referential essence of language. But it also points in a different direction. My criticism isn't the later Wittgensteinian critique that there are no significant a priori limits to the variety of uses of language, though that too has merit. My point is that there is no a priori requirement that representational thought be linguistic or symbolic. It is a further question whether there are significant a priori limits on the variety of different cognitions.

4.2. Atomic Propositions: Representation, Truth, and Individuation

The tractarian account of atomic propositions is an incomplete realization of three valuable insights. (i) Ordinary declarative sentences are representational, not because they stand in some relation to a primitively representational abstract object (a Fregean or a Russellian proposition), or because they name some bit of reality, but because of how they are used. (ii) Talk of these sentences being true or false is grounded in the fact that sentences are used to represent various things as bearing certain properties and standing in certain relations. (iii) The truth conditions of the use of an atomic sentence are read off the representational properties of that use—where a use is true at a world-state w iff were w actual, things would be as the use represents them.

This embryonic theory leaves it open that *different propositions* may be true at all the same world-states. It also leaves open other questions about the individuation of atomic propositions. Consider the proposition that Los Angeles is south of San Francisco. *What use of which sentence is it identical with?* Clearly, there is no more reason to identify it with a use of the sentence 'Los Angeles is south of San Francisco' than there is to identify it with a use of any other sentence that represents the same thing. Perhaps it should be something that all representationally identical uses of individual sentences have in common. Consider the act of using *some sentence or other* to represent Los Angeles as being south of San Francisco. Anyone who uses a sentence S in this way thereby performs the general representational act, but one can perform that general act without using S in particular. The general act is itself a proposition that *every agent* using an individual sentence in this way entertains.

Now go further. Consider the cognitive act of predicating *being south of San Francisco* of Los Angeles—i.e., of cognizing the two as being so related—*by any means whatsoever.* Surely, it is the best candidate for being *the proposition that Los Angeles is south of San Francisco*. If it's *not* possible perform it without using any sentence, then it is identical with the act of using some sentence or other to so represent the two cities. If it is possible to perform it without using any symbolic intermediary, then it alone is the proposition we seek. Either way, it is a proposition that anyone using any one of our atomic sentences entertains.

4.3. Truth-Functionally Complex Propositions

Suppose that atomic propositions are acts of representing objects as having properties, often or always by using sentences to do so. What about negations and disjunctions? Shouldn't they be acts of representing objects as being certain ways, where the objects represented are those their propositional constituents represent as being in various ways? We can bring this about by taking negation, disjunction, and the like to be operations we perform on propositions.

The negation of the proposition *that a is F*—which is the cognitive act of representing a as being F—can be taken to be the act of negating that proposition, which represents a as not being F. The disjunction of the propositions *that a is F* and *that b is G* may be identified with the act of operating on them to produce the proposition that represents the pair a,b as standing in the *two-place relation R* that consists of *the first's being F or the second's being G*. One who performs this act represents a and b as standing in this disjunctive relation, which is what it is to represent *a as being F or b as being G*. Applying negation to the disjunctive proposition generates the proposition that represents a,b as standing in the relation ~R that consists in *not being such that the first is F or the second is G,* or, more simply, *neither the first's being F nor the second's being G*. Other truth-functional operations are treated similarly, allowing us to say about them what the *Tractatus* says about atomic propositions: they represent tractarian objects as being certain ways, and so are true iff the objects are as they are represented to be. Both atomic propositions and truth-functional compounds represent objects as having properties that they possibly could have. Both are true iff the objects actually have those properties.[13]

This way of conceiving of truth-functionally compound propositions differs from one that takes them to predicate truth/falsity of their constituents. On that account, the negation of p predicates falsity (or perhaps *not being true*) of p*,* the conjunction of p and q predicates *being jointly true* of the pair, and the disjunction predicates *being true of at least one* of the pair. These truth-functionally compound propositions *directly represent* not tractarian simples as bearing properties or standing in relations, but simpler propositions as being true or false in various combinations. On this view, there is one theory of truth for both atomic and non-atomic propositions; a proposition is true if and only if things are as the proposition (directly) represents them to be. But now we allow not only objects in the world to be represented by virtue of being targets of predication, but also propositions about the world. However, *Tractatus* does not allow this.

[13] For more detail, including different but related ways of analyzing truth-functionally compound propositions, see Soames (2016).

The reason it doesn't begins with the question *What are we saying when we say that a proposition is true?* My answer has been *That things are as the proposition represents them to be*. Although this is the closest Wittgenstein comes to giving an unequivocal answer in the *Tractatus*, he isn't happy with it because he doesn't recognize the legitimacy of the question. As we shall see when the doctrine of *showing* is discussed in chapter 4, he thinks that *nothing* can be intelligibly *said* about the properties of propositions, the relations they bear to other propositions, or the relationship between propositions and the world (in virtue of which the former represent the latter). In part for this reason, he took a jaundiced view of *truth*, rejecting, in theory if not in practice, the idea that 'true' expresses a property that can be intelligibly predicated of anything. Thus, in the *Notebooks* (9(7)) he says that *'"p" is true' says the same thing as 'p'*. He would have been equally happy to say *the proposition p is true* is the very same proposition as *p*.

Although this sounds like it makes both forms of expression legitimate, that was not Wittgenstein's intention. Rather, he takes these predications of truth to be illegitimate. This is why he follows up the passage from the *Notebooks* with the remark that *'"p" is true'* is a pseudo-proposition, because it attempts to say something that can only be shown. Contrasting 'p' with ⌜'p' is true⌝—and implicitly with ⌜the proposition that p is true⌝—Max Black sums up the significance of this discussion for the *Tractatus* as follows:

> (a) [*'p' is true*] must be regarded as misleading and to be excluded from formulation in 'a correct ideography' [the ideal object language envisioned in the *Tractatus*]. For there is no place in Wittgenstein's conception of language for talk *about* propositions, as seems to occur in (a) [i.e., in *'p' is true*]. All significant propositions refer to the world by having their components stand proxy for objects in the world, but a proposition is not an object, and any method of symbolization that suggests the contrary must be incorrect. There is no room for a 'meta-language' in Wittgenstein's theory.[14]

Black is right. According to the *Tractatus,* (i) there can be no truth predicate of propositions, and (ii) there are no propositions that predicate any property or relation of propositions. Wittgenstein takes propositions to be facts that are logical pictures of other facts. Elementary propositions are combinations of names of metaphysically simple objects. Since propositions aren't metaphysical simples, there are no elementary propositions about propositions. Consequently, any propositions about propositions must be truth functions of elementary propositions about other things.

Now consider (i). Suppose for *reductio* that p is the proposition *that aRb*, and q is the proposition *that p is true*, which predicates *truth* of p. Since q isn't elementary, it must be a truth function of elementary propositions.

[14] Black (1964), p. 218.

Since elementary propositions are "logically" independent of one another, q can only be a truth function of p. According to Wittgenstein, however, p and q are consequences of each other. But then, by 5.141, q *is* the elementary proposition p, which merely predicates R of a and b.

> 5.141 If *p* follows from *q* and *q* follows from p then they are one and the same proposition.

Thus, q *doesn't* predicate truth of p. In short, there is no truth predicate of propositions.

Next, consider (ii). We know that if there are propositions that predicate anything of other propositions, they can't be elementary propositions, but must be truth functions of those. Let q be the proposition *that p is F*, where, for *reductio,* it is arbitrary what property is predicated of p. For there to be such a proposition q, the existence of p must be a truth-functional consequence of elementary propositions. Since p is a linguistic fact that represents a as bearing R to b, the existence of p requires the truth of the following claim C.

> Claim C: There are names a* and b* which, as a matter of linguistic convention, designate a and b respectively; a* stands in some structural relation R* to b* in p, and, as a matter of linguistic convention, for one name to stand in R* to a second name in a structure is for the structure to represent the object designated by the first name to stand in R to the object designated by the second name.

According to the Tractatus, it is impossible for Claim C to be a truth-functional consequence of elementary propositions because *there is no such complex proposition C at all*. This startling claim is a consequence of the tractarian doctrine that facts about the relationship between language and the world, in virtue of which (our use of) language represents the world, *cannot be stated in language*. Since I will discuss this paradoxical doctrine at length in chapter 4, I will here simply cite a few of the relevant passages in the *Tractatus* that articulate it.

At 2.18, we are told that logical form is what allows any picture, including any proposition, to represent the world.

> 2.18 What every picture, of whatever form, must have in common with reality in order to be able to represent it at all—rightly or falsely—is the logical form, that is, the form of reality.

What allows the proposition *that aRb* to represent reality as being a certain way is (i) that it contains names of a and b, and (ii) that the relation R* in which the names stand in the proposition represents the relation R in which it is possible for the objects designated by the names to stand. According to the *Tractatus*, neither (i) nor (ii) is capable of being stated in language.

That this is true of (i) is already implied by 3.202, 3.203, 3.36, and 3.263.

3.202 The simple signs employed in propositions are called names.
3.203 The name means the object. The object is its meaning.
3.26 The name cannot be analyzed further by any definition. It is a primitive sign.
3.263 The meanings of primitive signs can be explained by elucidations. Elucidations are propositions that contain the primitive signs. They can, therefore, only be understood when the meanings of these signs are already known.

At first glance, it appears to be a consequence of 3.203 that one can learn what a name—e.g., 'Tully'—means by learning which object it names, which, if one already understood 'Cicero', one could do if one were told "'Tully' names Cicero." But 3.263 denies this. For Wittgenstein, learning a name *isn't* learning a metalinguistic truth; it is coming to use the name to pick out the same individual that others do. Thus, it seems to be a consequence of the *Tractatus* that there are no genuine propositions stating the reference of names. A similar result holds for (ii) above.

These points are reinforced and summed up in the following passages.

4.12 Propositions can represent the whole of reality, but they cannot represent what they must have in common with reality in order to be able to represent it—logical form.
4.121 Propositions cannot represent the logical form: this mirrors itself in the propositions. [*Propositions cannot represent logical form: it is mirrored in them.*]
That which mirrors itself in language, language cannot represent. [*What finds its reflection in language, language cannot represent.*]
That which expresses *itself* in language, *we* cannot express by language. [*What expresses* itself *in language,* we *cannot express by means of language.*]
The propositions *show* the logical form of reality [*Propositions show the logical form of reality.*]
They exhibit it. [*They display it.*]
4.1211 Thus a proposition "*fa*" shows that in its sense the object *a* occurs, two propositions "*fa*" and "*ga*" that they are both about the same object. [*Thus one proposition* 'fa' *shows that the object a occurs in its sense, two propositions* 'fa' *and* 'ga' *show that the same object is mentioned in both of them.*]
If two proposition contradict one another, this is shown by their structure; similarly if one follows from another, etc. [*If two propositions contradict one another, then their structure shows it; the same is true if one of them follows from the other. And so on.*]
4.1212 What *can* be shown, *cannot* be said.

It follows that what is labeled 'Claim C' is not a genuine proposition. But if there is no such proposition, then there is no proposition q that predicates a property of the proposition *that aRb*, labeled 'p' above. In short, no

propositions predicate properties of propositions. This result is astounding.[15] According to the *Tractatus,* which says so much about propositions, nothing can be intelligibly said about them! Even if there is a way of retreating from this paradox, no coherent reconstruction of the *Tractatus* can take negations, conjunctions, and disjunctions to predicate truth or falsity of their constituent propositions.

The doctrine that one can't intelligibly predicate truth of propositions (as well as the doctrine that one can't intelligibly state the reference of names) is unfortunate, and can hardly be taken seriously by anyone wishing to give a semantic theory of referential uses of language or a philosophical theory of the nature of representational thought. Because Wittgenstein attempted to give us both, we are faced with two interpretative possibilities. One, suggested by Black, is to reconstrue some of his discussions of truth conditions and truth functionality, providing them with interpretations according to which truth is never predicated of sentences or propositions.[16] The other is to assiduously avert our eyes from the obviously incorrect doctrines about truth and reference until we are forced by the final few pages of the *Tractatus* to include them in the scope of the paradoxical tractarian conclusion that most of its central doctrines are unintelligible. My reading is a blend of these two strategies.

Propositions that predicate truth of other propositions can't be expunged from what is expressed by what are, in effect, sentences of the tractarian metalanguage. So, I will continue to say that negations are true whenever the negated propositions aren't true, and so on. But, if possible, one shouldn't interpret sentences of the imagined ideal tractarian object language as predicating truth, falsity, or anything else of propositions. This will, of course, limit its expressive power. For example, the object language described in the *Tractatus* should probably be taken to be, in principle, incapable of accommodating sentences used to report what agents believe, assert, or know. This, as we will see in chapter 4, was something Wittgenstein seems to have been prepared to accept. But it *must* include sentences expressing negative, conjunctive, and disjunctive propositions. Thus, we can't regard the negation of p as a proposition that predicates *being false*, or *not true*, of p. Nor can we take conjunctions or disjunctions of p and q to be propositions that predicate truth, falsity, or anything else of p and q.

My act-theoretic account of truth-functional compounds is consistent with this prohibition. The only other alternative I know of (which may, in fact, be Wittgenstein's) is mysterious. It simply asserts that the negation of the proposition p is *the unique proposition* that must be true if and only if p is not true—without explaining what that proposition is, what it

[15] The argument can be generalized to show that no propositions ever predicate properties of anything other than metaphysical simples.

[16] See the discussion of Wittgenstein's use of truth tables as propositional symbols on pp. 217–18 of Black (1964).

represents as being what ways, or *how it can have truth conditions at all*. The problem is repeated for the conjunction of p and q—which is defined as *the unique proposition* that must true if and only if p and q are jointly true—and for the disjunction of p and q—which is taken to be *the unique proposition* that must true if and only if either p is true or q is. You will be skeptical of these characterizations, if you recognize that necessarily equivalent propositions can fail to be identical. But Wittgenstein wasn't skeptical; he identified necessarily equivalent propositions. The mysterious analysis of truth-functional compounds requires this.

The analysis says nothing about negations, conjunctions, or disjunctions representing objects as being one way or another. So, *for them to be true can't be for them to represent objects as they really are*. Thus accepting the mysterious analysis requires positing two theories of truth—one defining truth for atomic *propositions* as representational accuracy and one defining truth for truth-functional compounds in terms of the truth or falsity of atomic propositions. Two theories of meaning are also needed. To know the meaning of an atomic *sentence* is to know which things it represents as being which ways. To know the meaning of a truth-functional compound is to know how its truth or falsity is determined by the truth or falsity of atomic sentences.

To this duplication, I add three related worries. First, if truth-functionally compound propositions can be identified only by using an illegitimate truth predicate, then no agent can identify them without affirming pseudo-propositions, and thereby making a mistake. That can't be right. Second, if *understanding* truth-functionally compound *sentences* requires *knowing* their truth conditions, which means knowing *they are true iff various atomic sentences or propositions are true (or false)*, then mastery of the "ideal" language of the *Tractatus* requires *knowing* pseudo-propositions. But that's impossible: *pseudo-propositions* can't be known. Third, any theory that identifies understanding some sentences with knowing their truth conditions must invoke a notion of truth in which the sentences S and ⌜'S' is true⌝ are *not* a priori consequences of one another. Wittgenstein had no such conception.

In short, Wittgenstein's text is inconsistent with any defensible account of *truth-functionally compound propositions*. There is, however, a defensible account that extends his insights about atomic propositions to truth-functional compounds. According to it, these propositions are acts of *using sentences* to represent tractarian metaphysical simples as having *properties* derived from those predicated by uses of atomic sentences. That *isn't* what Wittgenstein had in mind. But it preserves his most valuable insights.

4.4. General Propositions

Since general propositions—e.g., *that all Fs are G*—are central to the tractarian conception of logic that will be discussed in the next chapter, the point here will be preliminary. Wittgenstein treats general propositions as

a limiting case of truth-functional compounds with no upper bound on how many elementary propositions are needed to determine the truth or falsity of a general proposition. He expresses these propositions using his operator 'N' of joint denial, which takes indefinitely many propositions as arguments. Since the arguments can be given by complete sentences or by formulas containing free occurrences of variables, scope indicators he didn't provide are needed to indicate at what stage in the construction of a proposition variables get bound.

Here is an illustration.

THE TRACTARIAN TREATMENT OF *ALL F'S ARE G*.

1. *N(x[N(N(Fx), Gx)]* is true iff for every object o, a use of the formula *[N(N(Fx), Gx)]* in which 'x' is used to designate o is false.
2. A use of the formula *N(N(Fx), Gx)* in which 'x' designates o is false iff it is not the case that a use of the formula *N(Fx)* in which 'x' designates o is true and a use of the formula *Gx* in which 'x' designates o is false, i.e., iff it is not the case that o is F and o isn't G.
3. So *N(x[N(N(Fx), Gx)]* is true iff for every object o, it is not the case that o is F and o isn't G.

Consider all uses of *N(N(Fx), Gx)* in which 'x' is used to designate a metaphysical simple. Each use of the formula predicates *being both F and ~G* of the object 'x' is used to designate. The class of all such uses contains for each o, *the proposition that o is both F and ~G*. Applying joint denial to this class yields a proposition that denies each such proposition, and so, in effect, predicates the property of *not being F unless it is G* of each object. This will serve as the tractarian proposition that all Fs are Gs, if we can make sense of *indiscriminately predicating* a property of every object, including those we are not acquainted with.

As I have shown in Soames (2016), there is a natural way of doing this. Just as using a Fregean definite description when predicating *being so-and-so* of the object that satisfies the description amounts to predicating *determining something that is so-and-so* of the individual concept expressed by the description, so predicating *being so-and-so* of everything amounts to predicating *determining only so-and-so's* of a general concept that determines every object. Although this idea is not tractarian, it is a minimal modification that preserves the most important features of the neo-tractarian account of propositions developed here, while avoiding otherwise independent problems in the logic of the *Tractatus* to be taken up in chapter 3.[17]

[17] This way of understanding quantification exploits the fact that *unrestricted universal quantification* is the only quantification in the *Tractatus*. If the system included all generalized quantifiers—*all Fs, some Fs, most Fs,* and so on—we would be better off taking quantificational statements to predicate higher-order properties—e.g., *being true of all, some, or most Fs* – of lower-order properties.

4.5. The Second Fundamental Misstep: Identifying Equivalent Propositions

The central insight behind Wittgenstein's rejection of propositions as Frege, Russell, and Moore conceived them was that propositions are *not* abstract objects the representational nature of which, and truth conditions of, are independent of their role in the cognitive lives of agents. Instead, he rightly took their fundamental representational features to be, somehow, derived from agents' cognitions. This was an important advance. His starting point in articulating this idea was also insightful. Focusing on certain human artifacts—pictures, models, and sentences—he saw that our use of them to represent things as bearing properties and standing in relations was crucial to understanding propositions as pieces of information.

But there were problems in turning this promising starting point into a genuine solution to "the single great problem." The problems began with the error of running together a *sentence-as-used-to-represent-A-as-being-B* with *the use of a sentence to represent A as being B*. The latter is a cognitive act that represents the world because, necessarily, to perform that act is to represent the world. The former is a *pseudo-entity*: something that is somehow a contingent artifact—a structured combination of words and phrases—the representational features of which, and hence the truth conditions of which, are essential to it. There is no such thing. Whereas uses of sentences to perform specific representational acts have their truth conditions essentially, structured combinations of words don't.

The two ideas—uses versus sentences-as-used—also generalize differently. As shown in 4.2, the proposition *that Los Angeles is south of San Francisco* can't be identified with any single sentence, or with any *use* of a single sentence. It could, in principle, be identified with *a use* of any sentence to predicate *being south of San Francisco* of Los Angeles, or, even better, with the act of so predicating, with or without the help of a linguistic intermediary. In sections 4.3 and 4.4 I extended this idea by identifying compound propositions with uses of complex sentences to predicate complex properties of objects, or, even better, with acts of so predicating with or without the help of any sentence.

By contrast, it is hard to generalize the *sentence-as-used* idea. Starting with elementary propositions, one might identify the proposition that Los Angeles is south of San Francisco with the set of all sentences that mean that Los Angeles is south of San Francisco. But (i) this excludes indexical sentences like 'It is south of San Francisco' even though they can be used to say that Los Angeles is south of San Francisco. Since sentences have their meanings contingently, the proposal also leads to the unacceptable result (ii) that propositions may have different truth conditions at different world-states. Finally, the proposal identifies propositions not with facts but with sets. In addition to being inconsistent with the text of the *Tractatus*, it threatens the idea that propositions are inherently representational.

There is nothing about a set, no matter what its members, that represents anything as being any way. Hence, we don't think of them as having truth conditions.

On F. P. Ramsey's interpretation, Wittgenstein takes a slightly different tack in the *Tractatus*.[18] In effect, he posits the existence of highly abstract artifacts—*super-sentences* if you like—*instances* of which are either (a) abstract sentence types like 'Los Angeles is south of San Francisco', 'Los Angeles está al sur de San Francisco', and all other sentences sharing the sense of these two, or (b) sentence tokens that "have the same sense"—where to "have the same sense" is to represent the world in the same way.[19] Version (a) is subject to objections (i) and (ii) to the set-theoretic proposal just criticized. Version (b) may also be subject to objection (ii), if sentence tokens are taken to be sounds or visible marks produced by events of utterance or inscription. Since their representational significance depends on the linguistic conventions governing their production, it would seem that their truth conditions will vary from world-state to world-state. Version (b) also raises a worry about propositions no "tokens" of which have ever been produced (spoken, written, etc.), since in these cases the proposition types will be empty. Presumably, Wittgenstein wouldn't welcome the result that the propositions don't exist in such cases, nor would he welcome the result that two existent but empty types are identical, just as "two empty sets" are.

In the end, I am afraid that Ramsey's take on Wittgenstein's conception of propositions isn't specific enough to definitively evaluate. The use of the familiar type/token terminology is of little help because the sense in which it is here applied to propositions must be different from the antecedently understood sense in which we apply it to words, phrases, and sentences. For purposes of identifying propositions, one might as well have said that propositions are mysterious *we-know-not-whats* that are "expressed" by a bewildering variety of different sentence types or tokens at different world-states.

Another unclarity in Ramsey's characterization lies in filling out what it means for two sentences (types or tokens) to have the same sense (and hence to be instances of the same proposition). The most promising explication identifies the sense of a sentence with *the way it represents the world*. But, as Ramsey notes, the way in which this is most naturally understood applies only to atomic propositions.

> According to Mr. Wittgenstein a proposition token is a logical picture; and so its sense should be given by the definition of the sense of the picture;

[18] Ramsey (1923).

[19] Ramsey (1923) interprets the tractarian notion of propositions along the lines of (b). See p. 274 of the reprinting in Ramsey (1931).

> accordingly the sense of the proposition is that the things meant by its elements (the words) are combined with one another in the same way as are the elements themselves. . . . But it is evident that, to say the least, this definition is very incomplete; it can be applied literally only in one case, that of the completely analysed elementary proposition. . . . But this simple scheme must evidently be modified . . . if we have to deal with a more complicated proposition which contains logical constants such as 'not', or 'if', which do not represent objects as names do. . . . [This] difficulty must be faced, since we cannot be satisfied with a theory which deals only with elementary propositions.[20]

The tractarian way out is, he thinks, clear. As he puts it, "Mr. Wittgenstein says that any proposition is the expression of agreement and disagreement with the truth-possibilities of certain elementary propositions, and its sense is its agreement and disagreement with the possibilities of existence and non-existence of the corresponding atomic facts."[21]

It is helpful to spell out the tractarian idea of the sense of a sentence (type or token) as indicating agreement or disagreement with truth-possibilities in contemporary terms. Let S_1 and S_2 be any pair of sentences (types or tokens). Let w_i and w_j be any pair of possible world-states (the same or different) at which uses of S_1 and S_2 are governed by linguistic rules specifying what names in the sentences designate (when used at w_i/w_j), what structural relations in which the names stand represent the objects designated by them as standing in (when the sentences are used at w_i/w_j), and what expressions encode truth-functional operations at w_i/w_j. Given this, we derive results of the following form: S_1 *as used at* w_i is true-at-an arbitrary world-state w* if and only if at w* such-and-such is so-and-so; similarly for S_2 *as used at* w_j. This gives us the set $W\text{-}S_1w_i$ of world-states at which S_1 *as used at* w_i is true and the set $W\text{-}S_2w_j$ of world-states at which S_2 *as used at* w_j is true. We now say that the proposition p of which S_1 "is an instance or token" at w_i = the proposition q of which S_2 "is an instance or token" at w_j if and only if $W\text{-}S_1w_i = W\text{-}S_2w_j$.

Suppose that $W\text{-}S_1w_i$ is identical with $W\text{-}S_2w_j$. What, given this, is the single proposition p of which S_1 and S_2 are "instances or tokens" at the relevant world-states? Whatever it is, it's not something resembling a sentence, or set of sentences. There is no clear sense in which p is any linguistic artifact at all. Thus, we lose the initially promising thought that propositions are meaningful sentences, uses of sentences, or, more generally, artifacts put to certain representational uses. Since the "type"/"token" terminology is not helpful at this point, we might just as well retain the traditional terminology according to which propositions are *expressed* by sentence types or tokens. The most natural choice for the entity expressed by a sentence *as*

[20] Ramsey (1923) at pp. 275–76 of the reprinting.
[21] Ibid., p. 276.

used at a given world-state is then the set of possible world-states at which the sentence *as used at that state* is true. To maintain the spirit of the *Tractatus*, one might further stipulate that for an agent to entertain a proposition p (at w) is for the agent to use a sentence that expresses p (at w).

With this we derive one of the most important, but also most problematic, doctrines of the *Tractatus,* namely, that *necessarily equivalent propositions are identical.* In addition to recognizing only one necessary truth, the doctrine is inconsistent with the conjunction of what seem to be two obvious truths: (i) that one can believe or assert a proposition p without believing or asserting every necessary consequence q of p, and (ii) that one can't believe or assert a conjunction without believing or asserting both conjuncts.[22] (ii) is even inconsistent with the claim that one can believe or assert a necessary falsehood without thereby believing or asserting every proposition.[23]

My derivation of the doctrine, which relies on a theory of truth conditions of *sentences-as-used-at-a-world-state*, highlights a further difficulty. Although there are both sentences and uses of sentences (i.e., acts of using sentences to represent things as bearing properties and standing in relations), there are no such entities as *sentences-as-used-at-a- world-state*. There is, of course, a legitimate sense in which *sentences* can be assigned truth conditions in possible worlds semantics. What those who speak of S *as used at a world-state @ being true at w* are really saying is *that when S is used at @ it expresses a proposition that would be true were w actual*.[24]

With this we arrive at what may be the most revealing *reductio* of the tractarian conception of propositions. Wittgenstein's second fundamental misstep was to think he could abstract the notion of a proposition from the truth conditions of sentences. Informally put, propositions were to be what sentences with the same truth conditions have in common. Thinking of truth conditions in the way he implicitly did—as conditions in which sentences are true at world-states—leads, when spelled out in the detail we are now able provide, to the result just reached. A proper assignment of truth conditions to sentences at world-states presupposes *an antecedent conception of propositions*. Since this development of Ramsey's interpretation of the *Tractatus presupposes* propositions, it doesn't explain them.

With this in mind, return to the idea of *a particular type of use of a sentence* to, e.g., predicate a property or relation of an object or objects. Such a

[22] Suppose (a) that you believe or assert p and (b) that q is a necessary consequence of p. Since p is necessarily equivalent to the conjunction of p and q, the thesis that necessarily equivalent propositions are identical yields the result that you believe and assert the conjunction of p and q, which guarantees that you believe and assert q. See Soames (1987).

[23] This follows from the previous footnote plus the assumption that every proposition is a necessary consequence of any necessarily false proposition. For a more extensive discussion of the problems with identifying necessarily equivalent propositions, see chapter 3 of King, Soames, and Speaks (2014).

[24] See chapter 1 of Soames (2015b) for details.

use is true at a world-state w if and only if were the universe in state w things would be as the use represents them to be. What is *the way the use of the sentence represents things to be*? If the use is to predicate *being south of San Francisco* of Los Angeles, then the use represents Los Angeles as being south of San Francisco. If the use is to predicate *not being south of San Diego* of Los Angeles, then the use represents Los Angeles as not being south of San Diego. If the use is to predicate *being rational, if human* of every object, then the use represents every metaphysical simple as having that property. Crucially, what a use of a sentence represents is *not* indexed to a world-state. Remember, a use of a sentence is a type of cognitive act agents perform using the sentence. What the act represents is, by definition, what any actual or possible agent who used the sentence in that way would thereby represent—e.g., Los Angeles as being south of San Francisco, or Los Angeles as not being south of San Diego, or everything as being rational if human. Since this doesn't change from world-state to world-state, uses of sentences have their representational properties, and so their truth conditions, essentially.

This allows us to reconstruct a general account applying to all propositions that *vindicates* rather than *betrays* the promising insights that led Wittgenstein to his treatment of elementary propositions. We proceed in stages. At stage 1 we have propositions each of which is the act of using *a specific sentence* to predicate a property of objects. At stage 2 we have propositions each of which is the act of using *some sentence or other* to perform the predication. At stage 3 we have propositions each of which is the act of performing the predication *whether or not one uses any sentence to do so*. Each stage includes elementary and non-elementary propositions. At no stage is *truth at the same world-states* sufficient for propositions to be identical. At each stage, representing the same objects as bearing the same properties is necessary and sufficient for propositions to be *representationally identical*. If all that mattered was representational identity, genuine propositions could be limited to stage 3. If, more plausibly, fine-grained propositions are needed to deal with Frege's puzzle, then all three types should be recognized as genuine propositions.[25]

This analysis takes us well beyond the *Tractatus*, while capturing the insights behind its account of elementary propositions and avoiding the difficulties Wittgenstein had extending it to non-elementary propositions. It also avoids identifying necessarily equivalent propositions, which was a barrier to the breakthrough in our understanding language, mind, and information his account of propositions might otherwise have been. Of course, Wittgenstein himself would not have seen things this way. Without the identification of necessarily equivalent propositions, the *Tractatus*

[25] A much fuller account of the theory sketched here is given in chapters 2–5 of Soames (2015b).

would *not* have had the far-reaching consequences for philosophy, and its self-conception, that he passionately desired. These were the consequences that led him to take *the problem of the proposition* to be "the single great problem" of philosophy. Had he *correctly* conceived and solved that problem, he would have seen that its solution, though important to philosophy, linguistics, and psychology, wouldn't have been the world-changing event he dreamed of.

CHAPTER 3

The Logic of the *Tractatus*

1. TRUTH FUNCTIONALITY AND THE GENERAL FORM OF THE PROPOSITION

In chapter 2, I examined Wittgenstein's conception of propositions, according to which all propositions are truth functions of elementary propositions. This chapter explains how the *Tractatus* implements that idea. Since tractarian propositions are truth functions of elementary propositions, each non-elementary proposition p should be constructible by applying truth-functional operators to elementary propositions, collecting the results, and continuing to apply truth-functional operators until p is generated. Wittgenstein puts the point this way.

> 5. Propositions are truth functions of elementary propositions. (An elementary proposition is a truth function of itself.)
> 5.01 The elementary propositions are the truth-arguments of propositions.
> 5.3 All propositions are the results of truth-operations on the elementary propositions.
> The truth-operation is the way in which a truth function arises from [is produced out of] elementary propositions.
> According to the nature of truth-operations, in the same way as out of elementary propositions arise their truth-functions, from truth-functions

arises a new one. Every truth-operation creates from truth-functions of elementary propositions another truth-function of elementary propositions, i.e. a proposition. The result of every truth-operation on the results of truth-operations on elementary propositions is also the result of *one* truth-operation on propositions. [*It is of the essence of truth-operations that, just as elementary propositions yield a truth-function of themselves, so too in the same way truth-functions yield a further truth-function. When a truth-operation is applied to elementary propositions, it always generates another truth function of elementary propositions, another proposition. When a truth-operation is applied to the results of truth-operations on elementary propositions, there is always a single operation on elementary propositions that has the same result.*]

Every proposition is the result of truth-operations on elementary propositions.

5.5 Every truth-function is the result of the successive application of the operation (- - - - - T)(ζ,. . . .) to elementary propositions.

This operation denies [negates] all the propositions in the right-hand bracket and I call it the negation of these propositions.

Whereas standard logical systems have truth-functional operators '~', '&', '∨', '→', '↔', Wittgenstein had a single operator, 'N', for joint negation. Unlike the usual operators, which always attach either to a single sentence, as in ⌜~S⌝, or to a pair of sentences, as in ⌜(A&B)⌝, Wittgenstein's 'N' can apply to any number of sentences ⌜N(A)⌝, ⌜N(A,B)⌝, ⌜N(A,B,C)⌝ . . . to produce a complex sentence that is true if and only if all its argument sentences are false.

Wittgenstein's notation '(- - - - - T)' is unusual. The idea behind it can be illustrated using the truth tables for conjunction and disjunction.

A	B	A & B
T	T	T
T	F	F
F	T	F
F	F	F

A	B	A ∨ B
T	T	T
T	F	T
F	T	T
F	F	F

To construct a truth table of n arguments, one starts by assigning truth to each of the n elementary propositions and ends by assigning falsity to each, always proceeding in a fixed order (e.g., the assignment 'T, F' precedes the assignment 'F, T'). Given this order, one can present the two tables as '(T F F F)' and '(T T T F)', as Wittgenstein does at 5.101, or even as '(T - - -)' and '(T T T -)', which is the technique used for 'N' at 5.5. The remark at 5.3 that "When a truth-operation is applied to the results of truth-operations on elementary propositions, there is always a single operation on elementary propositions that has the same result" is illustrated below.

1a. (A & ~(B ∨ C)) ∨ ((~A & ~C) ∨ ((A & B) & C))

A	B	C	(A & ~(B ∨ C)) ∨ ((~A & ~C) ∨ ((A &B) & C))		
T	T	T	F	T	T
T	T	F	F	F	F
T	F	T	F	F	F
T	F	F	T	T	F
F	T	T	F	F	F
F	T		F	T	T
F	F	T	F	F	F
F	F	F	F	T	T

(1a) is a truth function of three arguments, constructed from elementary propositions by repeated application of '~', '&', and '∨'. Wittgenstein's point is that the same proposition results from one application of a single truth-functional operator— i.e., proposition (1a) = proposition (1b).

1b. (T F F T F T F T)(A,B,C)

The result can be reproduced whenever a proposition is a truth function of finitely many elementary propositions.

There is, of course, no point in introducing a new truth-functional operator for every truth function of n arguments—a total of 2 to the 2^n distinct n-place operators for each n— since each such operator can be defined using disjunction, conjunction, and negation. To do so in the case of (1a), we just read off, for each assignment of truth to the entire formula (i.e., the 1^{st}, 4^{th}, 6^{th}, and 8^{th} lines), a conjunction of elementaries or their negations, and then disjoin the conjunctions.

1c. (A & B & C) ∨ (A & ~B & ~C) ∨ (~A & B & ~C) ∨ (~A & ~B & ~C)

At 5.5, Wittgenstein takes this a step further, claiming that every truth-functionally compound proposition can be formulated as the result of successive applications of a single truth-functional operator, 'N', of joint denial. That this is correct for every truth-function of finitely many elementary propositions follows from the fact that (i) every such truth function can be defined using '~', '&', and '∨', and (ii) '~', '&', and '∨' are definable using 'N' as follows:

~P	N(P)
P & Q	N(N(P), N(Q))
P ∨ Q	N(N(P, Q))

Other equivalences include:

~P & ~Q	N(P, Q)
P & ~Q	N(N(P), Q)
~(P & ~Q)	N(N(N(P), Q))
~P ∨ Q	N(N(N(P), Q))
P → Q	N(N(N(P), Q))
P ↔ Q	N(N[N(P,Q)), N(N(P), N(Q))])

Next we turn to the claim that general propositions (which Russell and Frege expressed using quantifiers) are constructible from elementary propositions by successive applications of 'N'.

2. GENERALITY

Immediately after introducing his operator, 'N', Wittgenstein outlines how he will use it to express general propositions. In Russell's system, one starts with atomic formulas—e.g., 'Fa' and 'Gx'. Complex formulas are constructed in two ways: (i) by applying truth-functional operators—'~', '&', '∨', '→', '↔'—to get formulas like '(Ga & Hab) ∨ ~(Px → Qy)', and (ii) by applying existential and universal quantification to get sentences like '∃x Fx' and '∀x Fx'. For Russell, some sentences involve both sorts of complexity—e.g., '∀x (Fx → Gx)', which is constructed from the atomic formulas 'Fx' and 'Gx' by first using the truth-functional operator '→' and then adding the universal quantifier. The order in which the operations take place makes a difference. If we reversed the order, by first attaching the quantifier to the atomic formulas and then connecting the results with the truth-functional operator, we would get a different and non-equivalent sentence, '∀x Fx → ∀x Gx'. So, for Russell, complex sentences are built up from atomic formulas by finitely many applications of truth-functional and quantificational operators. Some compound sentences involve both kinds of operators, and the order in which they are applied makes a difference.

Wittgenstein intended his logical system to get essentially the same results as Russell by different means. Whereas Russell used quantifiers, Wittgenstein eliminated them.

> 5.521 *I dissociate the concept* all *from truth-functions.*
> *Frege and Russell introduced generality in association with logical product or logical sum. This made it difficult to understand the propositions '∃x Fx' and '∀x Fx', in which both ideas are embedded.*

Wittgenstein's idea was that the work of quantifiers be done by allowing 'N' to apply to *all* members of a specified class of propositions. He outlines the main idea in 5.501.

> 5.501 *When a bracketed expression has propositions as its terms—and the order of the terms inside the brackets is indifferent—then I indicate it by a sign of the form '(ξ̄)'. 'ξ' is a variable whose values are terms of the bracketed expression and the bar over the variable indicates that it is the representative of all its values in the brackets.*
> *(E. g. if ξ has three values P, Q, R, then*
>
> $$(\bar{\xi}) = (P, Q, R).)$$
>
> *What the values of the variable are is something that is stipulated.*

The stipulation is a description of the propositions that have the variable as their representative.

How the description of the terms of the bracketed expression is produced is not essential.

We can *distinguish three kinds of description: 1. direct enumeration, in which case we can simply substitute for the variable the constants that are its values; 2. giving a function* fx *whose values for all values of* x *are the propositions to be described; 3. giving a formal law that governs the construction of the propositions, in which case the bracketed expression has as its members all the terms of a series of forms.*

5.502 *So instead of '(- - - - - T)(ξ, . . .)' I write 'N(ξ̄)'.*

'N(ξ̄)' is the negation of all the values of the propositional variable ξ.

Generality is expressed by prefixing 'N' to a bracketed expression that represents all propositions of a certain sort. Sometimes the propositions to which 'N' applies are enumerated by listing them one by one. Sometimes they are given by "a function *fx*", which Wittgenstein takes to be a formula containing a variable. (The formula is often called "a propositional function.") In (2) of the final paragraph of 5.501, the variable 'x' in 'fx' ranges over individuals (metaphysical simples), so the propositions on which 'N' operates when prefixed to 'fx' includes all propositions that arise from the formula by replacing 'x' with a name of an individual. Since we don't know how many individuals there are, we don't know how many propositions 'N' operates on in such a case. What we do know is that it operates on *all of them*. That is the germ of the tractarian account of generality.

On this picture, the tractarian equivalent of ~∃*x Fx* is *N[Fx]*; its negation, *N(N[Fx]),* is equivalent to ∃*x Px*. Here, it is important to distinguish square braces, [ ], from round braces, ( ). Both are used in specifying the arguments of 'N'. But when variables are used to indicate generality, the use of square braces specifies the indefinitely large class of propositional arguments on which 'N' operates. Thus we must distinguish between *N(N[Fx])* and *N[N(Fx)]*. The latter is the joint negation of all propositions *N(Fa), N(Fb)*, *N(Fc)* . . . , which is equivalent to ∀*x Fx*. The former is the negation of *N[Fx]*—i.e., the negation of the joint denial of *Fa, Fb*, *Fc* . . . . Here are more examples.

~∀x Fx	N(N([N(Fx)]))
∃x ~Fx	N(N([N(Fx)]))
~∃x (Fx & Gx)	N([N(N(Fx), N(Gx))])
∃x (Fx & Gx)	N(N([N(N(Fx), N(Gx))]))
∃x (Fx & ~Gx)	N(N([N(N(Fx), Gx)]))
~∃x (Fx & ~Gx)	N([N(N(Fx), Gx)])
∀x ~(Fx & ~Gx)	N([N(N(Fx), Gx)])
∀x (Fx → Gx)	N([N(N(Fx), Gx)])

Proceeding in this way, we can express all sentences of the standard predicate calculus involving quantification of at most a single variable.[1] What we can't express, without more discriminating ways of specifying the arguments of 'N', are sentences involving mixed quantification like $\forall x \exists y\, Rxy$. For example, flat-footed application of the methods so far specified would result in the following equivalences:

N[Rxy]	$\sim\exists x \exists y\, Rxy$
N(N[Rxy]	$\exists x \exists y\, Rxy$
N[N(Rxy)]	$\forall x \forall y\, (Rxy)$
N(N[N(Rxy)])	$\sim\forall x \forall y\, (Rxy)$

Every application of 'N' to a class of arguments specified using different variables would have the (Russellian) effect of imposing the same binding conditions (universal or existential) on all the variables. To avoid this we must introduce a way of binding variables one at a time that Wittgenstein did not make explicit.

The idea that this is needed does, of course, presuppose that we need generality. It could be eliminated, if there were only *finitely* many objects (metaphysical simples) and we knew just which they were. But Wittgenstein denies that generality can be eliminated in this way.

> 4.1272 *Thus the variable name 'x' is the proper sign for the pseudo-concept* object.
> *Wherever the word 'object' ('thing', etc.) is correctly used, it is expressed in conceptual notation by a variable name.*
> *For example, in the proposition, 'There are 2 objects which . . .', it is expressed by '(∃x, y) . . .'.*
> *Wherever it is used in a different way, that is as a proper concept word, nonsensical pseudo-propositions are the result.*
> *So one cannot say, for example, 'There are objects', as one might say 'There are books'. And it is just as impossible to say, 'There are 100 objects', or, 'There are aleph-null objects'.*
> *And it is nonsensical to speak of the* total number of objects.
>
> 4.128(b) *[T]here are no preeminent numbers in logic, and hence there is no possibility of philosophical monism or dualism, etc.*

We can't, Wittgenstein tells us, make sense of any claim about how many objects there are because *object* is a formal, and hence not a genuine, concept or property. Although there *seems* to be a property that applies to all and only metaphysical simples, there can't be, because an elementary proposition predicating it of anything would be necessarily true or necessarily false, which no elementary proposition can be. Thus, there are no genuine propositions that say of things that they are objects. To think

[1] In relating tractarian to Russellian formulas, I have assumed the tractarian doctrines *that every object (metaphysical simple) over which we quantify has a tractarian name, that the same objects (metaphysical simples) exist at every world-state,* and *that names are rigid designators.*

otherwise is to imagine one can say, or state, what, the *Tractatus* tells us, can only be shown by our use of individual symbols. Thus, we must take tractarian generality to be non-eliminable.

The expressive incompleteness of our present understanding of generality may be summed up by saying that if there are infinitely many elementary propositions, then there are genuine tractarian propositions—i.e., truth functions of elementary propositions—that cannot be expressed in a version of the tractarian notation (described at 5.5, 5.501, and 5.502) in which the arguments of Wittgenstein's operator, 'N', are specified only by finite direct enumeration or by "a function [i.e., formula] *fx* whose values for all values of 'x' are the propositions described [i.e., to be operated on]."[2] Since Wittgenstein intended *all* propositions to be expressible, while declining to assume that there are only finitely many elementaries, he needed another, more flexible, means of specifying arguments of 'N'. The search for one begins with the final clause of 5.501, which tells us that the arguments of 'N' can be specified by "giving a formal law that governs the construction of the propositions, in which case the bracketed expression [that provides the arguments for 'N'] has as its members all the terms of a series of forms."

Wittgenstein explains what he means by "a series of forms" and his notation for it, which he calls "the general term of a series of forms," at 5.2521 and 5.2522.[3]

> 5.2521 *If an operation is applied repeatedly to its own results, I speak of successive applications of it. ("O'O'O'a" is the result of three successive applications of the operation "O'ξ" to "a".)*
> *In a similar sense I speak of successive applications of* more than one *operation to a number of propositions.*
>
> 5.2522 *Accordingly I use the sign "[a, x, O'x]" for the general term of the series of forms* a, O'a, O'O'a, . . . *This bracketed expression is a variable: the first term of the bracketed expression is the beginning of the series of forms, the second is the form of a term* x *arbitrarily selected from the series, and the third is the form of the term that immediately follows* x *in the series.*

We can, Wittgenstein thinks, specify an infinite series of terms using a "variable" of this sort. He uses this thought in giving the general term of a series of propositions, when, at 6 and 6.001, he explains what he means by "the general form of a proposition."[4]

> 6. *The general form of a truth function is* $[\overline{p}, \overline{\xi}, N(\overline{\xi})]$.
> *This is the general form of a proposition.*
>
> 6.001 *What this says is just that every proposition is a result of successive applications to elementary propositions of the operation* $N(\overline{\xi})$.

[2] 5.501(e).
[3] For more on operations and formal series, see 4.1252, 4.1273, 5.21–5.25.
[4] See also 5.234 and 5.3.

For Wittgenstein, the general form of something is analogous to a recursive definition of it. Such a definition of tractarian propositions could be expressed roughly as follows.

2a. That which is expressed by an n-place predicate followed by n names is an elementary proposition.
b. The result of applying the operation of joint denial N to any set of propositions is itself a proposition.
c. Nothing else is a proposition.

In (2) 'N' names a function that assigns, to any set of propositions as argument, a proposition as value that is the joint denial of the propositions in the selected set. So, (2) tells us that a proposition is either an elementary proposition, or what you get by (i) selecting any set of elementary propositions and applying N, (ii) collecting further propositions—including those arising from (i)—and applying N to any set of such propositions, and (iii) continuing the process without end.

Although this is illuminating, we still don't have what we need. In the tractarian object-language, 'N' is not the name of anything; it is a truth-functional operator. It is not a truth-functional operator in (2); it is the name of a function (the values of which are not truth values, but propositions). Since 'N' is part of a notation for formulating propositions, 'N(Pa)' is an instance of a proposition, not a name of one, which it is in (2). Moreover, in (2) the function N is not restricted to applying only to sets of propositions for which we have names. Even if a set of propositions has no name, (2) tells us that the result of applying N to it is a proposition. It may simply be that this proposition is not named by any singular term of the form 'N(. . .)'.

Nevertheless, (2) points us to what we need—an instruction telling us how each member of a series of sentences formulating every genuine proposition can be constructed by successive iterations of the operator $\mathcal{N}(\bar{\xi})$ prefixed to expressions representing arbitrary sets of propositions. This operator is not a function, but a *symbol* that attaches in the first instance to a *base expression*, representing an arbitrary set of elementary propositions. However, at this point there is a question the full significance of which Wittgenstein may not have appreciated: What expressions are available, first for representing arbitrary sets of elementary propositions, and then for representing arbitrary sets of propositions, whether elementary or not?

The question is related to a difference between the general term of a series of forms defined at 5.2521 and 5.2522 and the general form of the proposition given at 6 and 6.001. In the former, a formal series of expressions is generated from a single base term by an operation that applies to an arbitrary term to produce the unique succeeding term in the series. The result is a linearly ordered sequence of expressions. In the latter, the propositional series starts not with a single proposition, but with a selection from the set of elementary propositions. (The bar over 'p' at

6 indicates this plurality.) This is the initial stage in the series of stages. Later stages are generated from earlier stages by prefixing 'N' to expressions representing arbitrary sets of propositions drawn from earlier stages. Because each stage is potentially infinite, we don't get a linearly ordered sequence of propositions. Rather, we get a linear sequence of stages. Thus, Wittgenstein's claim that "*every proposition is a result of successive applications to elementary propositions of the operation* $N(\bar{\xi})$*"*($\bar{\xi}$) must be understood as asserting that every proposition is found at some stage.

Whether or not this claim is true depends on what expressions are available at each stage for representing arbitrary sets of propositions. What do we know about this? From 5.501 we know that finite enumeration is always available, as is replacing constants in sentences with variables. We also know from our discussion of generality that this is not enough. Fortunately, more is available. At 5.501 Wittgenstein also says, "How the description of the terms of the bracketed expression [representing arguments of 'N'] is produced is not essential." He is even more explicit at 3.317.

> 3.317 *To stipulate values for a propositional variable is* to give the propositions *whose common characteristic the variable is.*
> *The stipulation is a description of those propositions.*
> *The stipulation will therefore be concerned only with the symbols, not with the meaning.*
> *And the* only *thing essential to the stipulation is* that it is merely a description of symbols and states nothing about what is signified.
> *How the description of the propositions is produced is not essential.*

With this license to improvise, we are free to augment the notation given in the *Tractatus* for representing arbitrary sets of propositions as arguments of 'N', provided that we don't, in so doing, violate explicit tractarian doctrines.

It is useful to note that Wittgenstein anticipated the difficulties noted above involving multiple variables.

> 4.0411(a) *If, for example, we wanted to express what we now write as '∀x fx' by putting an affix in front of 'fx'—for instance by writing 'Gen fx'—it would not be adequate: we should not know what was being generalized. If we wanted to signalize it with an affix '*$_g$*'—for instance by writing '*$f(x_g)$*'—that would not be adequate either; we should not know the scope of the generality sign.*

The first of the problems noted here is that when two variables are present, we need a treatment of generality capable of generalizing on either one (existentially or universally) while leaving the other variable available for later generalization. It was because we couldn't do this in our earlier discussion that we couldn't express '∀x∃y Rxy'. The second problem noted at 4.0411(a) is that when '$f(x_g)$' is itself a constituent of a larger formula '. . . $f(x_g)$ . . .', it is unclear wither the scope of the generalizing operation is to be merely '$f(x_g)$'—resulting in something with the truth

conditions of '. . . $\forall x\, fx$ . . .'—or whether the scope of the operation is to be the larger formula—resulting in something with the truth conditions '$\forall x$ *(. . .* fx *. . .)*'. We met that difficulty by distinguishing square brackets from ordinary round brackets. What we need is a way of marking scope while simultaneously distinguishing generalization on one variable in a formula from generalization on another.

For this purpose we may introduce a tractarian language L_T that goes beyond Wittgenstein's explicit comments without violating tractarian strictures. Let an atomic formula be a predicate followed by n terms—i.e., names or (individual) variables. If $F_1 \ldots F_n$ are formulas of L_T and G is a formula of L_T in which the variable ν occurs free, then $\ulcorner(F_1 \ldots F_n)\urcorner$ and $\ulcorner(\nu[G])\urcorner$ are set representatives in L_T. The occurrence of ν to the left of G is a *generality indicator*. (Nothing else is a set representative.) If S is a set representative, $\ulcorner NS\urcorner$ is a formula of L_T. (There are no other formulas.). When a variable ν is used to form a set representative, it binds all occurrences of ν within G not already bound in G. Occurrences not bound in this way are free. A sentence is a formula with no free occurrences of variables. An atomic *sentence* is true if and only if its predicate applies to the objects named by its logically proper names. A *sentence* $\ulcorner NS\urcorner$ is true if and only if all *sentences* corresponding to the set representative S are false. If $S = \ulcorner(\nu[G])\urcorner$, then a sentence corresponds to S if and only if it arises from G by substituting occurrences of a single name for all free occurrences of ν in G. If $S = \ulcorner(F_1 \ldots F_n)\urcorner$, a *sentence* corresponds to S if and only if it is one of the Fs.

In the presence of our other tractarian assumptions, it can easily be shown by induction on the complexity of sentences that all propositions expressible in the first-order predicate calculus are expressible in L_T.[5] Examples are given below.[6]

~∃x Fx	N(x[Fx])
∃x Fx	N(N(x[Fx]))
~∃x ~Fx	N(x[N(Fx)])
∀x Fx	N(x[N(Fx)])
~∀x Fx	N(N(x[N(Fx)]))
∃x ~Fx	N(N(x[N(Fx)]))

[5] For a proof of essentially this result in a related system, see Schonfinkel (1924) at p. 358 of van Heijenoort (1967). The above system for reconstructing the logic of the *Tractatus* was presented in Soames (1983). A similar system can be found in Geach (1981). Some exchanges about these systems, pro and con, can be found in Geach (1982), Fogelin (1982), and chapter 6 of Fogelin (1987). The seeming expressive incompleteness of the explicit tractarian treatment of generality involving examples like '∀x∃y Rxy' was noted in Fogelin (1976).

[6] Sentences of the first-order calculus and their tractarian counterparts are logically equivalent in the sense of having the same truth values in every domain. However, the first-order calculus defines *logical truth* and *logical consequence* in terms of truth in all possible domains, no matter what size, while the tractarian presupposes a fixed domain. This leads to differences that will be explored later in this chapter.

~∃x (Fx & Gx)	N(x[N(N(Fx), N(Gx))])
∃x (Fx & Gx)	N(N(x[N(N(Fx), N(Gx))]))
∃x (Fx & ~Gx)	N(N(x[N(N(Fx), Gx)]))
~∃x (Fx & ~Gx)	N(x[N(N(Fx), Gx)])
∀x ~(Fx & ~Gx)	N(x[N(N(Fx), Gx)])
∀x (Fx → Gx)	N(x[N(N(Fx), Gx)])
∀y∃x (Rxy)	N(y[N(x[Rxy])])

The final equivalence in this list is established as follows: 'N(y[N(x[Rxy])])' is true iff each of the following is false: (i) 'N(x[Rxa])', (ii) 'N(x[Rxb])', (iii) 'N(x[Rxc])', and so on, one sentence for each object. That will be the case iff (i) '~∃x Rxa' is false, (ii) '~∃x Rxb' is false, (iii) '~∃x Rxc' is false, and so on, one of these statements for each object. That in turn will be the case iff (i) '∃x Rxa' is true, (ii) '∃x Rxb' is true, (iii) '∃x Rxc' is true, and so on, one such statement for each object. But that is the case iff for every object y it is true that ∃x Rxy—i.e., iff '∀y∃x (Rxy)' is true.

3. CARDINALITY, IDENTITY, AND EXPRESSIVE POWER

3.1. Infinity, the General Form of the Proposition, and the Predicate Calculus

We have seen that *if there are only finitely many tractarian objects* (metaphysical simples), then no quantifiers or symbols for generality are strictly required, because every tractarian proposition will be a truth function of *finitely many* elementary propositions. *If there are infinitely many tractarian objects,* then something like the treatment of generality in L_T is needed to vindicate the tractarian claim that every proposition is the result of successive application of the operator 'N'. But if there are infinitely many tractarian objects it might seem that there must be infinitely many tractarian names for them to ensure the equivalence of the tractarian system with standard versions of the first-order predicate calculus. That is worrisome, since presumably no human agent could master a language with infinitely many (primitive) logically proper names.

There is also another problem. Suppose there are infinitely many elementary propositions. Then, since Wittgenstein maintains that every set of elementary propositions is logically independent of every other set, it follows that every set of elementary propositions, finite or infinite, will determine a proposition the truth conditions of which differ from the truth conditions of every proposition determined by any other such set. This means, by Cantor's Theorem, that there will *uncountably many* different tractarian propositions.[7] But since there are only countably many

[7] Cantor's Theorem (published in 1891) demonstrated that for any collection C there is a collection of all subsets of C (often called 'the Power Set of C') that is strictly larger than C

expressions of our augmented tractarian language L_T (or of any language without infinitely long sentences), there will be infinitely many tractarian propositions that can't be expressed in the language. Thus, in the only sense we have been able to give to Wittgenstein's claim that "every proposition is a result of successive applications to elementary propositions of the operation N(ξ)," it would appear that the claim is false, if there are infinitely many tractarian objects, and infinitely many tractarian elementary propositions.

To what extent is the damage reparable? Although there is no limit on the size of domains of individuals over which sentences of the first-order predicate calculus are interpreted, models are generally required to interpret only finitely many names and predicates. So, there are only finitely many atomic sentences. Still, there may be infinitely many variables and hence infinitely many atomic formulas, all of which may play crucial roles in the assignment of truth conditions to quantified sentences. The trick is to assign truth conditions to these sentences that are derived, not from the truth conditions of atomic *sentences*, but from the truth conditions of atomic formulas *relative to assignments of objects to variables.* This was an innovation of Tarski (1935), which will be explained in chapter 9. Details aside, one can understand Tarski's idea as a way of treating variables, which are unbound in atomic formulas, as temporary names for objects when the truth conditions of such formulas relative to assignments are needed to evaluate quantified sentences. The extent to which this idea can be accommodated by the tractarian system is the extent to which the *Tractatus* can accommodate ordinary first-order quantification over infinitely many tractarian individuals without requiring infinitely many logically proper names.

Any attempt at accommodation must focus on the conjunction of two tractarian doctrines: (i) that elementary propositions consist of names standing in structural relationships to one another, and (ii) that every general proposition is both a truth function of elementary propositions and the result of successive applications of 'N' to propositions represented by the expressions that provide the arguments of 'N'. A natural, though flat-footed, idea would be to expand the class of elementary propositions to include structures obtainable from ordinary elementary propositions (structured combinations of names) by substituting variables for one or more of the names, and combining each resulting new structure with a second entity, a function mapping the variables onto objects. Although this might work technically, it is, philosophically, a nonstarter. One of the great achievements of the *Tractatus* was its insightful sketch—flawed and incomplete though it was—of a plausible, naturalistic theory of propositions. To

in the sense that it cannot be put in 1:1 correspondence with any subset of C (including C itself). See Soames (2014), pp. 269–70 for discussion.

sacrifice the promise of that attempt would be to betray a central part of its legacy.

There is a better alternative. In chapter 2, I argued that rather than viewing elementary propositions as bare syntactic structures, which Wittgenstein took to be facts of a certain sort, he would have done better to identify them with *uses of those structures*, which are acts of a certain sort. Let 'a' and 'b' be names and 'Rab' be the atomic sentence which, according to Wittgenstein, is the fact that consists of the symbol 'R' immediately followed by 'a', which is immediately followed by 'b'. Let 'Rxy' be the same, except that 'x' and 'y' are variables. Call anything that is a name or a variable a *term*. Now suppose it is a convention of the language that *any structure consisting of 'R' immediately followed by a term t_1 that immediately precedes a term t_2 is used to represent the object t_1 is used to designate as bearing the relation R* to the object t_2 is used to designate.* Let it be a further convention that speakers use 'a' and 'b' to designate objects o and o*, respectively. Now we need only suppose *that it is also a convention that an agent may use variables as temporary names for any object the agent wishes*. This gives us indefinitely many *uses of atomic formulas* to count as elementary propositions over and above the *uses of atomic sentences* that also count as elementary propositions. With this, we can accommodate first-order universal and existential generalizations without requiring a name for each object.

It is, of course, true that in order to achieve this result, we must countenance *uses of structures to represent this as bearing R to that* in cases in which the relevant structures are never, in fact, so used. But why should this be a problem? Everyone admits there are sentences that have never been uttered or inscribed, as well as complex syntactic structures that have never been the structures of any uttered or inscribed sentence. Well, *uses* are acts of a certain sort, and it is a commonplace that some acts that haven't been performed will be performed in the future, and that some acts that may never be performed, could be performed—including acts of using expressions in various ways.[8]

3.2. Higher-Order Quantification?

This result can be extended to give a tractarian treatment of second-order quantification into predicate position. There is nothing in the characterization at 5.501 of the ways in which arguments of 'N' can be specified that limits the treatment of generality to the expressive power of first-, as opposed to higher-order, quantification. Indeed, at 3.317, 4.0411, and 5.501 we have already seen a hint that second-order quantification into predicate position may be possible. One standard treatment of such quantification

[8] The legitimacy of quantifying over the merely possible is defended at pp. 128–29 of Soames (2010a). See also Soames (2007a) and Salmon (1987).

takes second-order quantifiers to range over all subsets of the domain of the first-order quantifiers.[9] So, if there are infinitely many individuals in the domain, there will be uncountably many sets of individuals in the range of such quantifiers. Although it had not been established at the time the *Tractatus* was written, it is now well known that this second-order quantification increases the expressive power of a language, while rendering sound proof procedures for the resulting formal system incapable of deriving all logical truths, as well as all logical consequences of sentences in the system.[10] The question here is whether the tractarian system can be extended to include second-order quantification.

I don't see any very plausible way of doing so. In addition to incorporating the suggestion made in the previous section that propositions be identified not with sentences but with *uses* of *sentences and formulas,* we could, of course, extend L_T to L_{T2} by adding n-place second-order variables ranging over n-place relations on individuals. As with tractarian names, we need not impose limits on how many predicate constants may be needed. As with tractarian simples, we could assume that the same relations on simples are available at every possible world-state. We need not require all n-place relations on simples be named by predicate constants, and we could recognize uses of predicate variables to represent uncountably many relations on n-tuples of simples. Whether or not Wittgenstein would have wanted this expressive power is another matter. He could have had it. However, it would have been of doubtful utility for him.

The atomic formulas of L_{T2} are syntactic structures in which an occurrence of either an n-place predicate constant or an n-place predicate variable (over n-place relations) is followed by n occurrences of singular terms, which are either proper names or first-order variables (over individuals). Let $F_1 \ldots F_n$ be formulas of L_{T2}, G be a formula of L_{T2} in which the first-order variable ν occurs free, and H be a formula of L_{T2} in which the second-order variable V^n occurs free. Then $\ulcorner(F_1 \ldots F_n)\urcorner$, $\ulcorner(\nu[G])\urcorner$, and $\ulcorner(V^n[H])\urcorner$ are *set representatives*. If S is a set representative, then $\ulcorner NS\urcorner$ is a formula. When a variable ν is used to form a set representative, it binds all free occurrences of ν in the formula to which it attaches; similarly for second-order variables. Occurrences not bound are free. A sentence is a formula with no free occurrences of variables. For each name n there is a convention stipulating that n is used to designate a certain specific object. For each of n-place predicate constant P^n there is a convention that a structure consisting of P^n immediately followed by terms $t_1 \ldots t_n$ represents the objects those terms are used to designate as standing in a certain specific relation R. There is also a convention that speakers can use any

[9] The relationship between first-order and higher-order quantification in Frege's system is explained on pp. 25–26 of Soames (2014). Similar ground concerning Russell's system is covered on pp. 500–504.

[10] See Soames (2014), pp. 26–29, and also pp. 504–7; pp. 511–24 are also relevant.

individual variable to designate any individual, named or unnamed, and any n-place relation variable V^n to represent any n-place relation on individuals, whether represented by a predicate constant or not. (To say that one uses a predicate variable V^n to represent R is to say that the speaker uses structures in which V^n is followed by n occurrences of singular terms to represent the objects those terms are used to designate as standing in R.) These conventions are presupposed in specifying what it is for a *use* of a sentence or formula of L_{T2} to be true.

A use (consistent with the conventions of L_{T2}) of an atomic *sentence* or *formula*—consisting of an n-place predicate constant or variable P^n followed by n occurrences of singular terms—is true if and only if the objects the terms are used to designate stand in the relation R represented by the use of P^n. A use u (consistent with the conventions of L_{T2}) of a *sentence* $\ulcorner NS \urcorner$ in which S is a set representative is true if and only if (i) $S = \ulcorner(F_1 \ldots F_n)\urcorner$ and for each F_i, the sub-use of F_i that is part of u is false, or (ii) $S = \ulcorner(\nu[G])\urcorner$ and for each object o, a use of G that involves letting ν designate o while agreeing with u on all other expressions is false, or (iii) $S = \ulcorner(V^n[H])\urcorner$ and for each n-place relation R on individuals, a use of H that involves letting V^n represent R, while agreeing with u on all other expressions, is false.

Moving from sentences and formulas to their uses provides us with uncountably many *uses* of formulas involving individual or predicate variables. As before, we think of them abstractly, countenancing *uses of formulas to represent things as being various way* in cases in which those formulas are never, in fact, used in those ways. This allows uses of sentences of the ideal tractarian object language to express previously unexpressed propositions that are truth functions of elementary propositions. This extension of expressive power is significant, even though it doesn't allow uses of sentences of L_{T2} to express of every proposition represented by a member of the power set of elementary propositions.

It is not clear how well the extension we have achieved fits into the *Tractatus* as a whole. For one thing, the account of second-order quantification brings with it an explicit ontology of n-place relations, which, if countenanced by the *Tractatus*, must, along with particular objects, be *metaphysical simples*. Whether or not Wittgenstein wished them to be included is difficult to determine. Although it is possible to read him as allowing it, I am not sure that the *Tractatus* settles the matter. Thus, I am inclined to share Black's flexible opinion.

> [Objects] are the materials of which atomic facts are constructed, the substance of the world (2.021). And they have form (2.025). We may think of an object's form . . . as manifested in restrictions upon the set of objects with which it can combine to produce atomic facts.
>
> Wittgenstein's view of 'objects' is very schematic. His conviction that propositions have a definite sense . . . drives him to postulate that there must be simples. . . . But about the logical form of these objects he has nothing definite

> to say. It would certainly be a mistake to identify objects with what we commonly call 'individuals', or to suppose that they cannot be at all like what we commonly call 'relations'. Since objects constitute the substance of the world, it is natural to think of them as timeless (cf. 2.207) and so to imagine them as resembling 'universals' rather than 'particulars', but both of these traditional terms are inappropriate. All we can really know about objects is that they exist.[11]

If this is right, then the ontology of relations required by our tractarian reconstruction of second-order quantification is neither explicitly tractarian nor definitely beyond the pale.

It may be more problematic that the identification of elementary propositions with *uses* of atomic sentences and formulas challenges the independence of elementary propositions, and, derivatively, of atomic facts. Consider sentences (3a,b), their respective second-order logical consequences (4a,b), and the corresponding atomic formulas (5a,b).

3a. Plato was a philosopher & Aristotle was a philosopher & ~Pericles was a philosopher.
b. Plato was a philosopher & Aristotle was a philosopher.
4a. $\exists V^3$ (V^3 Plato, Aristotle, Pericles)
b. $\exists V^2$ (V^2 Plato, Aristotle)
5a. V^3 Plato, Aristotle, Pericles
b. V^2 Plato, Aristotle

Next consider the elementary propositions that are identified with the uses of these atomic formulas specified by (6a) and (6b).[12]

6a. the use of (5a) to represent Plato, Aristotle, and Pericles as standing in the relation that R^3 that requires its first two arguments to be philosophers and its last argument not to be a philosopher.
6b. the use of (5b) to represent Plato and Aristotle as standing in the relation R^2 that requires its first two arguments to be philosophers.

Since it is impossible for (6a) to be true without (6b) being true, these two elementary propositions are not independent, nor are the corresponding atomic facts that make them true.

The difficulty uncovered here involves the following features of our extension of the tractarian system.

A. The tractarian independence of elementary propositions, and of atomic facts
B. The analysis of general propositions as truth functions of elementary propositions

[11] Black (1964), p. 57.
[12] Recall that propositions are uses of sentences or formulas in accordance with conventions. This is such a use.

C. The need to accommodate infinite domains of individuals, and hence (in the presence of B) the existence of uncountably many propositions, and of general propositions that are truth functions of collections of elementary propositions that together involve infinitely many individuals
D. The seeming intelligibility of higher-order quantification and hence the need (in the presence of B and C) of tractarian propositions, generalizing over relations, which are truth functions of uncountably many elementary propositions
E. The collapse of metaphysical and epistemological modalities into logical modalities

As we have seen, it is hard to jointly maintain all of these.

The problem arises from running together sentences with propositions, which, in turn, facilitates the confusion of logical modalities with metaphysical and epistemic modalities. The atomic *sentences* (excluding identities) of any standard logical system are independent of one another, in the sense that none is a logical consequence of any others. This is so because logical consequence, as we now understand it, is defined as truth preservation across all models. Since models are free to *reinterpret* all nonlogical vocabulary—i.e., all vocabulary appearing in atomic sentences—these sentences can't stand in logical relationships that preserve truth across all models. By itself this tells us nothing about whether the propositions expressed by particular uses or interpretations of atomic sentences are conceptually, or metaphysically, consistent with one another, or stand in relations of necessary or a priori consequence.

Since the conception of logic on which this criticism is based was not current when Wittgenstein wrote the *Tractatus*, and would not become so for more than a decade, he can't be blamed for not adhering to it. Still, his distinction between propositions and propositional signs plus his recognition of the role of conventions governing uses of language in making that distinction was a promising beginning. As the critique here illustrates, that beginning could have led to identifying propositions with *uses* of both sentences and formulas, thereby increasing the expressive power of the tractarian account of logic and language. However, this, and the unraveling of the independence doctrines for elementary propositions and atomic facts, would, as the following sections will make clear, have been only the beginning.

3.3. The Proper Understanding of Generality

Consider the claim that the proposition *that every object is F*—⌜∀x Φx⌝ in Russellian notation—is expressible in a tractarian system using ⌜N(x[N(Φx)])⌝. As we have seen, the latter is the result of negating every member of the set S of all propositions *that o isn't F,* for each object o—each such proposition being expressed by a use of ⌜N(Φx)⌝ in which 'x' designates o and the

structure consisting of Φ followed by 'x' represents o as being F. This set S is given via the set representative ⌜(x[N(Φx)])⌝. But how is it given? Is it *named*, in which case the semantic content of ⌜(x[N(Φx)])⌝ (contributed to the general proposition) is simply the set S? Or is S *described*, in which case the semantic content of ⌜(x[N(Φx)])⌝ is (something like) the property *being a set of propositions consisting, for each object o, of a use of* ⌜*N(Φx)*⌝ *in which 'x' designates o and the structure consisting of* Φ *followed by 'x' represents o as being F*? There are problems either way.

Suppose S is named. Then, the tractarian general proposition is a truth function of the set S_E each member of which is the elementary proposition *that o is F* expressed by a use of ⌜(Φx)⌝ in which 'x' designates o. So, not surprisingly, S_E and the proposition expressed by ⌜N(x[N(Φx)])⌝ are a priori consequences of each other. But this is *not* true of the proposition *that every object o is F.* Although that proposition *is* true if and only if each proposition in S_E is true, it is possible, no matter whether S_E is finite or infinite, for an agent to *know* each of its members, and even to know that every proposition in S_E is true, without knowing whether or not there are more propositions *that o is F* that are not members of S_E, and so without knowing, or being in a position to come to know by a priori reasoning, *that every object is F.* Thus, the general proposition is *not* an a priori consequence of S_E. Nor, it appears, are the individual members of S_E a priori consequences of the general proposition. For example, I can know that *that every man is mortal,* without knowing, for each man, a singular proposition about that man.

If this argument is correct, then the understanding of ⌜N(x[N(Φx)])⌝ according to which the propositions provided by ⌜(x[N(Φx)])⌝ are directly named rather than described *fails* to express the proposition, *that every object is F*, which is expressed in standard notional by ⌜∀x Φx⌝. No similar argument applies to the understanding of ⌜N(x[N(Φx)])⌝ according to which the propositions provided by ⌜(x[N(Φx)])⌝ are *described* in the manner previously indicated. There is, however, a different worry. Since the description of the arguments of 'N' speaks of them as *representing each object o as being F,* it violates the tractarian proscription against speaking of the representational relationship between propositions and the world. Since this difficult doctrine will be taken up in the next chapter, I will put it aside for now. Assuming, for now, that the descriptive understanding is legitimate, we note that on this understanding, ⌜N(x[N(Φx)])⌝ is *not* an a priori consequence of S_E and the individual propositions in S_E are *not* a priori consequences of the proposition expressed by ⌜N(x[N(Φx)])⌝. This is as it should be, even though each *sentence* ⌜Φn⌝ may still properly be regarded as a *logical consequence* of ⌜N(x[N(Φx)])⌝ or ⌜∀x Φx⌝ in any system in which a *sentence* Q is a logical consequence of a *sentence* P if and only if, for every model M that interprets both, if P is true in M, so is Q. One could even preserve the idea that ⌜N(x[N(Φx)])⌝ and ⌜∀x Φx⌝ are logical consequences of some set of atomic *sentences* if one restricted the models

to those with a fixed cardinality and required every object to be named. However, there is little interest, outside the *Tractatus,* in such a restriction.

Indeed, if one takes all tractarian restrictions seriously—including the requirements (i) that all objects are named by rigid designators, (ii) that the same objects exist at every possible world-state, (iii) that all and only necessary truths are a priori truths, and (iv) that all and only a priori truths are logical truths—then all the distinctions made in this section are obliterated. Since the distinctions are clearly significant, they provide further information about why some tractarian restrictions can't be accepted, while raising the question of how much can be saved of Wittgenstein's account of generality, if they aren't.

The lesson here is that rejecting the restrictions while trying to retain the skeleton of the tractarian account of generality will push one a long way toward the familiar analysis of quantification as higher-order predication descending from Frege and Russell.[13] On that analysis, the proposition *that every object is F* doesn't predicate anything of individual objects, individual propositions, or sets of propositions. Instead, it predicates the higher-order property *being true of every object*, of the property *being F.* To incorporate this idea into the tractarian treatment of generality would be to take a proposition expressed by a use of ⌜N(x[N(Φx)])⌝ (at the actual world-state) to predicate something like *being a property of a set of propositions that contains only untruths* of the property *being a set consisting of, for each (existing) object o, the negation of the proposition that o is F.* This, or some variant, will accommodate the points made here about a priori and logical consequence, and extend them to necessary consequence, when contentious tractarian assumptions are relaxed. However, one may doubt whether what amounts to a predicate of properties of sets of propositions can properly be regarded as a joint negation operator on propositions—in which case one may suspect that all that has been saved of the tractarian analysis of generality is its form.

3.4. The Tractarian Attack on Identity

Generality, which is standardly expressed in logic by the universal and existential quantifiers, is not the only logical notion Wittgenstein sought to improve on. He also had problems with identity, standardly expressed by '='. As noted in chapter 1, he couldn't take identity to be a relation on objects, nor could he take '=' to be a predicate appearing in elementary propositions.

(i) If identity were a relation on objects, then for each object o, there would be *a fact* consisting of o's being combined with o in the requisite way. But

[13] See Soames (2014), chapters 1, 2, and 8.

what means of combination is that? If we try to think of such a fact, all we end up thinking of is o itself, which, it would seem, is an object, not a fact.

(ii) If identity were a relation on objects and there were a convention to use '=' to represent objects as standing in that relation, then there would be elementary propositions expressed by uses of 'a = b', 'b = c', and 'a = c'. But these propositions are *not* logically independent of each other. Hence, there can be no such propositions.

In addition to these problems, which arise from metaphysical and linguistic doctrines of the *Tractatus*, there is a deeper worry (iiia), which is exacerbated in (iiib) by the tractarian collapse of the metaphysical and epistemic modalities into the logical modalities.

(iiia) If identity were a relation on objects, then to say of o that it is identical with o would be to say something trivial and uninformative, while to say of some distinct o* that it is identical with o would be say something too obviously false to ever say.

(iiib) If identity were a relation on objects, then to say of o that it is identical with o would be to assert a necessary a priori truth, with no cognitive significance, while to say of two different objects that they are identical would be to assert a necessary a priori falsehood, which, by Wittgenstein's criterion of propositional identity, is a senseless contradiction.

Although one can understand Wittgenstein's concern over (i) and (ii), they need not trouble those who don't subscribe to tractarian doctrines about atomic facts and elementary propositions. But (iiia), which is a version of Frege's puzzle, is genuinely problematic, especially for one like Wittgenstein, who, rightly, rejected Frege's proposed solution, which involved distinguishing the meanings of names from their referents.[14] Nor is the point behind (iiib) easily dismissible, even if one rejects Wittgenstein's attempt to reduce both metaphysical and epistemic modalities to logical modalities. Since the proposition *that o is identical to o* is both necessary and knowable a priori, it is natural to think that when o isn't identical to o* *the proposition that o ≠ o** is also both necessary and knowable a priori, *in which case the truth or falsity of every elementary proposition involving identity is knowable a priori*. If that were so, one might certainly question whether such propositions were ever worth asserting or denying.

All of this is puzzling enough. Things become more puzzling when one notices that many thoughts we express using the identity predicate seem to be perfectly significant. What are we to make of this? Are there really no such significant thoughts? Are they all, unbeknownst to us, really nonsense, or are they genuine thoughts that need expressing in some, perhaps other, way? Wittgenstein addresses these points in the following passages.

[14] For explication and criticism of Frege's puzzle and his proposed solution, see Soames (2014), pp. 86–96.

5.53 Identity of the object I express by identity of the sign and not by means of a sign of identity. Difference of the objects by difference of the signs.

5.5301 That identity is not a relation between objects is obvious.

5.5303 Roughly speaking: to say of *two* things that they are identical is nonsense, and to say of *one* thing that it is identical with itself is to say nothing.

5.531 I write therefore not "F(a,b) & a = b", but "F(a,a)" (or "F(b,b)"). And not "F(a,b) & ~(a = b)", but "F(a,b)".

5.532 And analogously: not "(∃x,y) [F(x,y) & x = y]", but "(∃x) F(x,x)"; and not "(∃x,y) [F(x,y) & ~(x = y)]", but "(∃x,y) F(x,y)".

5.5321 Instead of "∀x (Fx → x = a)" we therefore write e.g. "[(∃x) Fx → (Fa & ~(∃x,y) (Fx & Fy))]". And the proposition "*only* one x satisfies F( )" reads: "[(∃x) Fx & ~(∃x,y) (Fx & Fy)]".

5.533 The identity sign is therefore not an essential constituent of logical notation.

5.534 And we see that apparent propositions like: "a = a", "(a = b & b = c) → a = c", "∀x (x = x)", "∃x (x = a)", etc. cannot be written in a correct logical notation at all.

5.535 So all problems disappear which are connected with such pseudo-propositions.

The ideas expressed in these passages are a mixture of the unremarkable—5.53, 5.531, 5.532, 5.5321—and the astounding—5.5301, 5.5303, 5.534, and 5.535. The former illustrate a notational proposal for expressing propositions without the identity sign that are truth-conditionally equivalent to propositions normally expressed with it. The latter provide a general statement of Wittgenstein's proposal and attempt to explain why it is philosophically required. This is where things become truly puzzling. *Both the articulation of the proposal in 5.53 and the statement of the rationale for it in the next two passages use the very notion they repudiate as unintelligible.* But if identity makes no sense, how are we supposed to understand Wittgenstein's proposal, or to know how to implement it?

Consider 5.53. It tells us that *for all objects o_1 and o_2 Wittgenstein will express the claim that o_1 is identical with o_1 by using a single name, and he will express the claim that o_1 is not identical with o_2 by using non-identical names.* But if the claim that *that a is, or isn't, identical with b* is a mere pseudo-proposition, as alleged at 5.535, then the claim announcing Wittgenstein's proposal is also a pseudo-proposition. How, then, can it be informative—if the notion required to understand it makes no sense? The same point can be made about attempts to implement the proposal, or to assess whether Wittgenstein follows it in practice. To do either we must know, for various expressions e_1 and e_2, whether or not e_1 is identical with e_2, while also knowing, of the objects o_1 and o_2 named by a pair of expressions, whether or not they are identical. In short, if, as we are told, identity makes no sense, then Wittgenstein hasn't introduced any alternative; if, on the other hand,

identity does make sense, then we have no need for his notational alternative, even though we can understand and evaluate whether we would lose anything by adopting it.

Since we can't give up identity, we had better address the puzzles that led Wittgenstein to reject it. The key tractarian passage is 5.5303, which combines (iiia) and (iiib) above. The former, (iiia), is essentially Frege's puzzle for Millianism about names—the doctrine that the meaning (semantic content) of a name is its referent—and the corollary that if n and m are two names of a single object o, then the proposition expressed by a use of ⌜n = m⌝ is the trivial proposition *that o is identical with o*. This puzzle is challenging because the bare proposition *that o is identical with o* is necessary, knowable a priori, and, seemingly, uninformative, or, empty. Given this, one might well wonder why we ever need to express it. The classical Fregean response *denies* that the proposition expressed by ⌜n = m⌝ is the bare proposition *that o = o*. Instead, Frege takes it to be an abstract combination of the different meanings of n, m, and of '=' (whatever they may be). Wittgenstein rightly rejects this mysterious entity. Not seeing an alternative, he was led to the present impasse. What he didn't know, and is even now not widely known, is that there are plausible versions of Millianism that escape the problem.

In fact, one such version is already implicit in the *Tractatus*-inspired analysis of propositions introduced in chapter 2. The analysis identifies *some propositions* with *uses of sentences* (or formulas) to represent things as bearing various properties and relations, while identifying other propositions as similar acts of representation, abstracting away from which, if any, sentences are used. With this in mind, compare P1–P3.

P1. The cognitive act of using n to pick out o, m to pick out o, and ⌜n = m⌝ to represent the objects so named as being identical.
P2. The cognitive act of using n to pick out o and ⌜n = n⌝ to represent o as being identical with o.
P3. The act of representing o as being identical with o, however o is picked out and whatever sentence, if any, is used.

Since one can perform the first of these acts without performing the second, proposition P1 is different from proposition P2. Since anyone who performs either of these acts thereby performs the third, but not conversely, P3 is different from both P1 and P2. It will then follow that anyone who entertains, asserts, believes, or knows either P1 or P2 thereby entertains, asserts, believes, or knows P3—but not conversely.

Next we take advantage of a commonplace about the names used in everyday life. One can successfully use each member of a pair of different names—'Mark Twain' and 'Samuel Clemens', 'Cicero' and 'Tully', 'Hesperus' and 'Phosphorus', 'London' and 'Londres', 'Peking' and 'Beijing', etc.—to designate the same object without knowing that the names designate the same thing. Applying this lesson to m and n, we get the result that

entertaining, asserting, believing, or knowing P2 and P3 is not sufficient for entertaining, asserting, believing, or knowing P1. So, whereas, P2 and P3 are knowable a priori (because there are ways of entertaining them for which no empirical knowledge is needed to determine their truth), P1 is not knowable a priori. Since P1 is informative in ways that P2 and P3 are not, to assert P1 is *not* to say something too obvious to be worth saying. Nor, if n* and n designate different objects, is the assertion made using ⌜n* = n⌝ epistemically equivalent to the assertion of a contradiction, or to the assertion of any other obvious falsehood. All of this is so, despite the fact that P1, P2, and P3 represent the very same thing as being the very same way, and so have identical truth conditions.[15]

In this way one may dispose of the objection (iiia) to the identity predicate, voiced at 5.303. To do so, however, one must disregard Wittgenstein's denial of an assumption just invoked—namely, that one can understand two codesignative names without knowing them to be codesignative. Russell also denied this, when the names were what he called "logically proper," which, he reasoned, could be used to refer only to those entities—oneself and one's private sense data—about which one couldn't be mistaken.[16] Since Wittgenstein's metaphysical simples are not easily conceived of as private sense data, it is surprising that he thinks that one who uses two names to refer to the same thing must *always* know that they do. But he does.

> 4.243 Can we understand two names without knowing whether they signify the same thing or two different things? Can we understand a proposition in which two names occur, without knowing if they mean the same or different things?
>
> If I know the meaning of an English and a synonymous German word, it is impossible for me not to know that they are synonymous, it is impossible for me not to be able to translate them into one another.
>
> Expressions like "a=a", or expressions deduced from these are neither elementary propositions nor otherwise significant signs. (This will be shown later.)

Since we now know that the second paragraph in this passage is false, we shouldn't be inhibited by it in proposing a solution to Frege's puzzle that disposes of Wittgenstein's key objection to the identity predicate.[17] How-

[15] Although the use of different expressions is crucial to distinguishing P1 from P2, neither proposition predicates anything of expressions, or represents them as having any properties. Hence, the truth conditions of the three propositions are identical. For detailed presentation and discussion of this point, see chapter 4 of Soames (2015b).

[16] For explanation and criticism of Russell's views on this point, see Soames (2014), pp. 386–88, 395–400.

[17] Saul Kripke (1979, 1980), Reiber (1992), Salmon (1986, 1989, 1990); Soames (1986), Soames (2002), pp. 67–72, and Soames (2015b), chapters 4 and 9.

ever, it is important to be aware that his belief in this falsehood was one of the reasons he did not see our proposed solution.

Having dealt with the objection (iiia) to the claim that identity is a relation on objects (expressed by the identity predicate), it remains to fully dispose of (iiib). Let o_1 and o_2 be distinct objects, let n and m be two names for o_1, let r name o_2, and let P1–P3 be as above. Finally, let P1~ and P3~ be as follows.

P1~ The cognitive act of using n to pick out o_1, r to pick out o_2, and ⌜n ≠ r⌝ to represent the objects so named as *not* being identical.

P3~ The cognitive act of representing o_1 as *not* being identical with o_2, however the two objects are picked out and whatever sentence, if any, is used.

Then, all five propositions are necessary truths, but only P2 and P3 are knowable a priori. P1 and P1~ are not knowable a priori because knowing them to be true requires empirical information about what the names refer to. P3~ fails to be knowable a priori because there is *no way* of entertaining it for which empirical evidence isn't required to determine its truth.[18] All of this would, of course, have been foreign to Wittgenstein, telling as it does against his collapsing of epistemic and metaphysical modalities. But it does help us more fully understand how and why his discussion of identity ended up in a *cul-de-sac*.

Having reinstated identity, we can evaluate his notational proposal, now understood not as a way of eliminating a problematic notion, but as an alternative way of securing the benefits of a useful one. When the proposal is understood in this way, it is easy to identify its shortcomings. Suppose that Wittgenstein's suggestion is correct: for every truth that can be expressed using '=', there is a truth-conditionally equivalent proposition expressed in a system without '=' in which different names always designate different objects (and similarly for uses of different variables). This is *not* sufficient to vindicate Wittgenstein's proposal. *What must be shown is that for every sentence $S_=$ containing '=' which an agent A knows he or she could use to express a proposition p, there is an alternative sentence S_W without '=' that A knows that he or she could use in accord with Wittgenstein's notational rule to express a proposition q that is truth-conditionally equivalent to p.* This can't be shown, because it isn't true. (Assume that propositions are truth-conditionally equivalent iff they are true at the same possible world-states.)

Suppose I don't know whether the names 'm' and 'n' (rigidly) designate the same object, but I do know I can use (7) to express a true proposition p.

7. Fn & Gm & (~(n = m) → Rnm)

I know that p is necessarily equivalent to the proposition $p_=$ that I could assert using (8a) *if 'm' and 'n' are codesignative*, while also knowing that p is

[18] See pp. 375–76 of Soames (2003a).

necessarily equivalent to the proposition $p_{\neq}$ that I could use (8b) to assert *if 'm' and 'n' designate different things.*

8a. Fn & Gn
b. Fn & Gm & Rnm

But I don't use either sentence to assert $p_{=}$ or $p_{\neq}$ because I don't know whether or not the names designate the same thing. I do know that, *if it is possible to use (8b*) in accord with tractarian conventions,* then such a use would assert a proposition necessarily equivalent to $p_{\neq}$.

8b*. Fn & Gm & Rnm

But this does me no good. Since I don't know whether or not 'm' and 'n' designate different things, I don't know whether I can use (8b*) in accord with the tractarian convention. Thus, I don't know how to express in tractarian notation the knowledge I know I can express using (7).

I do, of course, know I can express that knowledge without employing '=' *by using (9) in accord with the ordinary, non-tractarian, notational convention*.

9. (Fn & Gn) ∨ (Fn & Gm & Rnm)

But I don't know that I can use (9) *in accord with the tractarian convention,* because to know that I would have to know that 'n' and 'm' designate different objects, which I don't. Hence, the tractarian proposal leaves no way of knowing how to express the knowledge I wish to express.

The point can be underlined by comparing the conjunction, (10a), of (9) and 'Fm', *understood in the tractarian way*, with the conjunction, (10b), of (7) and 'Fm', *understood in the ordinary, non-tractarian way.*

10a. Fm & [(Fn & Gn) ∨ (Fn & Gm & Rnm)]
b. Fm & [Fn & Gm & (~(n = m) → Rnm)]

Whereas the truth of the *tractarian* (10a) requires the existence of two Fs, the truth of the *non-tractarian* (10b) does *not* require the existence of two Fs. This could only be so if the proposition expressed by a use of (9) *understood in the tractarian way* is epistemically more demanding than the proposition expressed by a use of the non-tractarian (7). The reason I can't use (9) in accord with the tractarian convention to express the knowledge I use (7) to express, is that I don't know the proposition expressed by such a use of (9), though I do know the less demanding proposition I use (7) to express. Hence Wittgenstein's notational replacement for the identity predicate was as inadequate as were his reasons for wanting a replacement.

3.5. Identity, Tautology, and Modal Collapse

One of the interesting uses of the identity predicate is in constructing, for every natural number n greater than 1, a sentence of the first-order

predicate calculus that is true in all and only interpretations with at least n objects.

11. $\exists x \exists y (x \neq y), \exists x \exists y \exists z (x \neq y \,\&\, x \neq z \,\&\, y \neq z), \ldots$

Since the domain of quantification can be any size, no sentence on this list is a logical truth. Were the tractarian system L_T to contain an identity predicate, we could construct a similar list starting with (12), which is true if and only if there are at least two objects.

12. N (N (x [N (N (y [N (x = y)])])))

Would any sentences on such a list express tractarian tautologies? To make sense of the question, while maintaining the doctrine of the independence of elementary propositions, we would have to treat '=' as a special *logical* symbol, while excluding propositions containing it from counting as *elementary*. Suppose we did. We could then ask, *Do any propositions in a list starting with (12) come out true on all assignments of truth values to elementary propositions?* The answer is, *Of course they do*. Suppose there are at least two metaphysical simples, which, according to the *Tractatus*, exist at all possible world-states. No matter what truth values are assigned to genuine elementary propositions predicating properties and relations of simples, the fact that there are *two* simples will ensure that the first item on the list is true. Hence it is a tractarian tautology. If there are infinitely many metaphysical simples, then all members of the list are tractarian tautologies. If there are exactly n simples, then the first n – 1 members are tractarian tautologies, and the rest are tractarian contradictions. But then since, for Wittgenstein, tautologies are truths that are both necessary and knowable a priori, it will follow that, for every proposition on the list, either it, or its negation, is an a priori, necessary truth.

Since this result is a *reductio ad absurdum* of the ideas generating it, Wittgenstein had to avoid it. It may seem that he did, despite the other shortcomings of his proposal to replace the identity predicate by the notational convention discussed in the previous section. In fact, he didn't. Given the convention that different names, and different variables, stand for different things, one can construct tractarian propositions that generate the same *reductio ad absurdum*.

13a. $\exists x \exists y [(Ax \vee \sim Ax) \,\&\, (Ay \vee \sim Ay)], \exists x \exists y \exists z [(Ax \vee \sim Ax) \,\&\, (Ay \vee \sim Ay) \,\&\, (Az \vee \sim Az)], \ldots$

b. $\exists x \exists y (R^2xy \vee \sim R^2xy), \exists x \exists y \exists z (R^3xyz \vee \sim R^3xyz), \ldots$

The first sentence on each list is true at all tractarian world-states at which at least two simples exist. The second sentence is true in all such states at which at least three simples exist, and so on, as in (11). The same lists could be repeated in tractarian notation. Since Wittgenstein hasn't avoided the *reductio*, both his identification of necessity and apriority with logical

necessity (i.e., tautology) and his doctrine that logical necessity is determinable by form alone are threatened.

He could, of course, consider giving up not only '=' but also his notational convention. Since this would leave many intelligible thoughts inexpressible, it is not a happy result for an analysis that purports to encompass every intelligible proposition. But this is only the beginning. Such a policy would also wreck the extension, in section 3.2 above, of the tractarian system to encompass the second-order predicate calculus. To see this, consider the second-order sentence (14), which is true in any model in which some subset of the domain of individuals contains one of those individuals while failing to contain some individual in the domain.

14. $\exists P \exists x \exists y\ (Px \,\&\, \sim Py)$

This sentence is true in all models that contain at least two individuals. The result generalizes, giving us, for each natural number n, a second-order sentence true in all models with at least n individuals. We get the same thing in our extension L_{T2} of the tractarian system, which leads to disastrous results when coupled with Wittgenstein's doctrine that the number and identity of metaphysical simples remains fixed across world-states.

In section 3.2 I showed that L_{T2} allowed uses of predicate variables occurring in atomic formulas to designate arbitrary subsets of the domain of individuals to count as elementary propositions, which, in turn, required relaxing the doctrine that elementary propositions be independent of one another. What I didn't point out then was that coupling this relaxation with adherence to other tractarian doctrines would require constraining possible assignments of truth values to elementary propositions. If the doctrines about metaphysical simples are retained, some constraints will arise from their number. If there are at least two simples, then any possible assignment must assign different truth values to some pair of elementary propositions, conceived of as uses of 'Px' in which the variable 'P' picks out the same set, and the uses of 'x' pick out different simples. Such constraints will yield tractarian "tautologies" that are counterparts of (14). Avoiding this reinstatement of the *reductio* requires abandoning L_{T2}.

A similar result will be reached in any system that allows what are now called "generalized quantifiers," including 'all Fs', 'most Fs', 'some F', and 'at least n F's'. Surely sentences containing these quantifiers are intelligible and are used to express genuine propositions. Although there is no bar to expressing them in extensions of the tractarian systems exhibited in this chapter, they will generate unwanted tractarian "tautologies" unless radical adjustments are made in Wittgenstein's underlying assumptions—among them, the independence of elementary propositions, and of atomic facts, the analysis of general propositions as truth functions of elementary propositions, and the collapse of metaphysical and epistemological modalities into logical modalities.

4. WITTGENSTEIN'S GENERAL LOGICAL DOCTRINES

The treatment of generality and identity in the *Tractatus* had little historical impact on the subsequent development of logic. The tractarian idea that logic is the study of *propositions*, and the way they represent the world, was shortly to give way to the modern conception, ushered in by Gödel and Tarski, of logic as the model-theoretic study of guaranteed truth preservation among *sentences* across systematic reinterpretations of nonlogical vocabulary. The fact that we have now traveled so far down this latter road is one of the chief obstacles faced by the modern reader who wishes to understand the impact of the *Tractatus*. That impact came, not from the fine points of tractarian logic we have been considering, but from the sweeping lessons about thought and language Wittgenstein drew from his analysis of the proposition. The most important of these were L1–L3.

L1 All necessity and apriority is linguistic necessity, and so the result of our system of representing the world, rather than the world itself. There are propositions that are necessarily true and knowable a priori, but there are no necessary facts to which they correspond. Rather their necessity and apriority is due to the meanings of words.
L2 All linguistic necessity is logical necessity.
L3 All logical necessity is determinable by form alone.

The most significant components of the tractarian analysis of the proposition that Wittgenstein took to support L1–L3 were the following.

A. The independence of elementary propositions (and of atomic facts)
B. The doctrine that all propositions are truth functions of elementary propositions, and indeed are the results of successive applications of the single truth-functional operator 'N'
C. The doctrine that propositions are abstract linguistic types the instances of which are truth-conditionally equivalent sentences, which means that truth-conditionally equivalent propositions are identical
D. The doctrine that elementary propositions—which are structures that represent objects as standing in various relations—are true if and only if there are atomic facts that consist in objects standing in the relations in which they are represented as standing
E. The doctrine that the truth of a non-elementary proposition doesn't consist in its correspondence with a non-elementary fact, because there are none, but rather is determined by the truth values of elementary propositions

The general picture that emerges from (A–E) is that elementary propositions and their negations are representations of how things are, or are not, in the world, while non-elementary propositions are merely summaries of elementary propositions and their negations. With the exception of tautologies and contradictions, their truth or falsity is simply a matter of which atomic facts there are. Since contradictions and tautologies don't

say anything significant, all thought and talk that is both significant and fully intelligible is reduced to humdrum attempts to report atomic facts in the world.

But surely, one is inclined to object, there is more. In addition to truths about how things are, which are known empirically, there *are* truths about how things *must be*, which are known a priori. "How can that be?" Wittgenstein would reply. Think about elementary propositions P and Q, which predicate unanalyzable properties of objects. What can we make of the claim that Q is an a priori consequence of P? If we simply have a pair of unanalyzable properties plus two bare sequences of objects, what could possibly explain a deductive inference from P to Q? When the question is put this way, one can understand the attraction of Wittgenstein's answer. Nothing, it would seem, could explain that inference! Thus, one might concede, no *elementary propositions* are a priori consequences of other elementaries, and none are inconsistent with others. But if *all propositions* are merely truth functions of elementary propositions, then, surely, *all a priori consequences* must be *logical consequences,* and all a priori inconsistencies must be *logical inconsistencies*. A similar result may seem to reduce necessary consequence and inconsistency to logical consequence and inconsistency. Thus, Wittgenstein's modal collapse, wrong though it may be, appeared explanatorily hardheaded, and even attractive.

Having come this far, we can complete the tractarian defense of L1 and L2 by recalling basic assumptions about the relation between elementary propositions and possible world-states (i.e., ways the universe could be, or could have been).

(i) Elementary propositions are true at some possible world-states and false at others.
(ii) Each elementary proposition is independent of all others; it is possible for it to be true (or to be false) no matter what truth values the others have.
(iii) A possible world-state is nothing over and above a collection of possible atomic facts.

It follows that there is a one-to-one correspondence between possible world-states and assignments of truth values to elementary propositions. In this way, the metaphysical modalities are reduced to logical modalities.

A proposition p is necessary—i.e., it would have been true no matter which possible state the universe had been in—if and only if p is logically necessary (a tautology).
A proposition p is impossible—i.e., it would have been false no matter which possible state the universe had been in—if and only if p is logically impossible (a contradiction).
A proposition p is contingent—i.e., it would have been true had the universe been in certain possible states, but false had the universe been in other states—if and only if p is neither a contradiction nor a tautology.

According to Wittgenstein, logically necessary propositions and logically impossible propositions are degenerate propositions. Consider tautologies. Since they are true at all world-states, they don't tell us anything about the actual state of the world that distinguishes it from any other state. In that sense they don't *say* anything. Rather, they are simply the result of having a symbol system that includes truth-functional operators.

> 6.1 The propositions of logic are tautologies.
> 6.11 The propositions of logic therefore say nothing. (They are analytical propositions.)
> 6.111 Theories which make propositions of logic appear substantial are always false. [*All theories that make a proposition of logic appear to have content are false.*]

We use truth-functional operators to say that *the world is not so and so*, and *the world is either such and such or so and so*. But once we have the operators, tautologies result from combining them in certain admissible ways. So, the thought goes, tautologies are nothing more than artifacts of our symbol system. When we recognize that $\ulcorner(A \vee \sim A)\urcorner$ and $\ulcorner((A \,\&\, (A \rightarrow B)) \rightarrow B)\urcorner$ are tautologies, we don't grasp metaphysically necessary facts; we simply see something about how our symbolism works. For example, we see that our conventions dictate that B follows from A, and $\ulcorner(A \rightarrow B)\urcorner$. Of course, the tautology doesn't say *that B follows from A and* $\ulcorner(A \rightarrow B)\urcorner$. Rather, it *shows* that without *saying* anything.[19]

Since tautologies are products of the symbolism, it may seem natural to suppose that one can always tell whether a proposition is a tautology just by examining how it is symbolized. Wittgenstein tells that one can always do this.

> 6.113 It is the characteristic mark of logical propositions that one can perceive in the symbol alone that they are true; and this fact contains in itself the whole philosophy of logic. And so also it is one of the most important facts that the truth or falsehood of non-logical propositions can *not* be recognized from the propositions alone.
> 6.126 Whether a proposition belongs to logic can always be calculated by calculating the logical properties of the *symbol*.
> And this we do when we prove a logical proposition. For without troubling ourselves about a sense and a meaning, we form the logical propositions out of others by mere *symbolic rules*.
> 6.127 Every tautology itself shows that it is a tautology.

This brings us to the Tractarian doctrine L3 that logical necessity is always determinable by form alone. It has two natural interpretations.

[19] See 6.1201.

Wittgenstein clearly meant (at least) that there is *a sound, complete, effective, positive test* for tautology—one which, given any tautology as input, will *always* tell us, in a finite number of steps, that the input proposition *is* a tautology, and will never wrongly classify, as a tautology, any input proposition that isn't a tautology (though it may sometimes yield no result for such an input). Simply put, Wittgenstein thought there were formal proof procedures one can use to prove all and only tautologies. But he may well have meant something stronger. He may have thought that there is an effective *decision procedure* which, when applied to any proposition, will *always* correctly tell us, in a finite number of steps, *whether or not it is a tautology*. (Such a procedure combines an effective positive test for tautology with an effective negative test.) Wittgenstein seems to suggest this when at 6.126 he says that we can always recognize *whether* a proposition is a tautology.

It would have been natural for him to think this, since his model, the propositional calculus, is decidable in this sense. In it, every proposition is either elementary, or the result of *finitely* many applications of truth-functional operators to *finitely* many propositional arguments. Because of this limitation, the truth-table method, illustrated in section 1, is a decision procedure for the system. Given a proposition constructed from n elementary propositions, one creates a table representing each of the 2^n possible assignments of truth and falsity to them. For each such assignment, one computes the truth or falsity of the entire proposition. If all these calculations yield truth, the proposition is a tautology; otherwise it isn't.

The logical system in the *Tractatus* is like the propositional calculus in some ways and unlike it in others. It is like the propositional calculus in that every proposition is taken to be a truth function of elementary propositions. It is unlike the calculus in allowing propositions that are truth functions of infinitely many propositions. This isn't, in itself, decisive, however. There is a decision procedure for tautology for certain limited systems incorporating generality by essentially tractarian means, even though they allow propositions that are truth functions of infinitely many elementary propositions.[20] Because of the way in which generality is constrained in these systems, their propositions can be arranged in a two-level list—the first consisting of infinitely many elementary propositions and the second consisting of a linear sequence of non-elementary propositions in which each proposition on the list is constructed by prefixing 'N' to an expression representing earlier propositions on the two-level list. This linear sequence makes a decision procedure possible.

Wittgenstein may have been thinking along these lines. At 6.1203, he sketches an elaborate procedure, applicable to propositions in which "no

[20] See the radically incomplete system described in Soames (1983).

sign of generality occurs," for deciding whether or not any given proposition is a tautology. This is followed at 6.122 by a sweeping pronouncement, *not* limited to those that contain "no sign of generality," about the possibility of eliminating "logical propositions," i.e., *all tautologies*—which would seem to imply the ability to recognize just which propositions are to be eliminated and which are to remain.

> 6.122 It follows from this that we can actually do without logical propositions; for in a suitable notation we can in fact recognize the formal properties of propositions by mere inspection of the propositions themselves.

This is unfortunate. In section 2, I constructed an essentially tractarian logical system L_T incorporating generality that can be given an interpretation in which it is expressively equivalent to the standard first-order predicate calculus. As we will see in chapter 8, a little over a decade after the *Tractatus* was published, the mathematician and philosopher Alonzo Church proved that *no decision procedure* for determining logical truth is possible for standard versions of the first-order predicate calculus. That result applies to L_T, in which some propositions involve a potential *infinity of applications of the truth-functional operator 'N'* to other propositions. For example, 'N(y[N(x[Rxy])])' arises from applying 'N' to each of the potential infinity of propositions: (i) 'N(x[Rxa])', (ii) 'N(x[Rxb])', (iii) 'N(x[Rxc])', ... The same is true for each member of that series—e.g., 'N(x[Rxa])' arises from applying 'N' to each of the potential infinity of propositions 'Raa', 'Rba', 'Rca', 'Rda', ... Thus, 'N(y[N(x[Rxy])]' arises from applying 'N' to a potential infinity of propositions each of which arises from applying 'N' to a potential infinity of propositions. There is no truth table for this and, if we allow infinite domains of objects, no decision procedure for logical truth.

However, since, on this understanding, L_T is equivalent to the first-order predicate calculus, there is a sound, complete, effective positive test for logical truth (tautology) in L_T.[21] Although the matter is disputed, it is, I think, plausible that Wittgenstein wished his system to have at least this expressive power.[22] Thus, the best interpretation of his doctrine L3 may simply be that there is such a positive test for tautology. But even this is problematic. Wittgenstein might have wished his tractarian system to have the expressive power of L_{T2}, developed in section 3.2. Apart from the other problems we noted, there is a natural way of interpreting L_{T2} in which it is equivalent to the standard second-order predicate calculus—for which, we now know, there can be no complete, effective, positive test for logical truth (tautology).[23] Finally, there is the matter, investigated in

[21] This result about the predicate calculus, which is due to Gödel, is mentioned in chapter 8.
[22] See chapter 6 of Fogelin (1987) for a dissenting opinion.
[23] See chapter 8.

3.5, of whether the tractarian system inadvertently gives us unwanted tractarian tautologies based on the number of metaphysical simples. For all these reasons, the question of whether there is an interpretation of Wittgenstein's doctrine L3 that is both acceptable and in accord with his intentions is vexed.

CHAPTER 4

The Tractarian Test of Intelligibility and Its Consequences

1. The Intelligibility Test
2. The Limits of Intelligibility: Value, the Meaning of Life, and Philosophy

1. THE INTELLIGIBILITY TEST

Chapter 3 closed with a discussion of difficulties with Wittgenstein's identification of necessity and apriority with logical necessity, discoverable by an examination of logical form alone. According to the *Tractatus*, every intelligible proposition p falls into one or the other of two categories: either (i) p is contingent (true at some possible world-states and false at others), in which case p is both a truth-function of elementary propositions and something that can be known to be true or false only by empirical investigation, or (ii) p is a tautology or contradiction that can be known to be so by formal calculation. The paradigmatic cases of meaningful uses of language for Wittgenstein are those in the first category. The uses in the second category are deemed meaningful because they are the inevitable product of the rules governing the logical vocabulary used in expressing the propositions of the first category. For Wittgenstein, tautologies and contradictions are uses of sentences that don't state anything, or give any information about the world. But their truth or falsity can be calculated, and understanding those uses reveals something about the symbols involved. Thus, the sentences, and their uses, can be regarded as intelligible in an extended sense.

Many uses of language that purport to make statements don't fit neatly into either category. Chief among them are attempts to state fundamental claims of ethics, aesthetics, and traditional philosophy. Since the sentences used for these purposes typically purport to state necessary truths that don't seem to be capable of being known on the basis of empirical observation, they seem *not* to fit into Wittgenstein's first category. Since they

don't seem to be tautologous or contradictory statements, the truth or falsity of which can be determined simply by examining linguistic form, they don't seem to fit into his second category. Because his doctrine purports to state conditions that must be satisfied in order for any use of a sentence to make a statement, he concludes that there are no genuine propositions of ethics, aesthetics, or traditional philosophy, and that the sentences used in these domains are nonsensical; they fail to be meaningful even in the extended sense in which tautologies and contradictions are. Thus we have what seems to be a powerful intelligibility test that categorizes masses of apparently meaningful uses of language as nonsensical.

There are, however, two difficulties extracting consequences from it. First, Wittgenstein never gives examples of metaphysical simples or elementary propositions. Despite maintaining that these mysterious entities must exist in order for any of our talk to make sense, central tractarian doctrines make it all but impossible to specify any. This makes it difficult to apply the intelligibility test. Since no elementary propositions are identified, whether claims made in science and everyday life are truth functions of them is problematic. How are we supposed to decide whether the claims *that uranium atoms are unstable, that space is curved, that heat is molecular motion,* and *that other minds exist* satisfy the condition, if we don't know which propositions are elementary?

The second difficulty is that we often can't apply the intelligibility test unless we know the logical form of a sentence. According to Wittgenstein, however, the logical forms of propositions expressed by uses of sentences of ordinary language are hidden, and revealed only by analysis. This is indicated at 4.002, where he elaborates on the hiddenness of logical form, and the difficulty of providing analyses.

> 4.002 *Man possesses the ability to construct languages capable of expressing every sense, without having any idea how each word has meaning or what its meaning is—just as people speak without knowing how the individual sounds are produced.*
>
> *Everyday language is part of the human organism and is no less complicated than it.*
>
> *It is not humanly possible to gather immediately from it what the logic of language is.*
>
> *Language disguises thought. So much so, that from the outward form of the clothing it is impossible to infer the form of the thought beneath it, because the outward form of the clothing is not designed to reveal the form of the body, but for entirely different purposes.*
>
> *The tacit conventions on which the understanding of everyday language depends are enormously complicated.*

This doctrine of hiddenness greatly inhibits our ability to apply the intelligibility test. If logical form is hidden, then, when confronted with a use of a sentence one suspects must be necessary, if the sentence is meaningful

at all, one may not know how to determine whether the necessity of the putative proposition is discoverable from the logical form of the sentence alone. We know from the test that if necessity can't be determined from form alone, then the sentence is nonsense and its use does not count as a genuine proposition. However, since logical form is hidden, we may not be able to apply the test. This difficulty may not arise in every case, but it does arise in some, and it is always in the background. Thus, Wittgenstein's intelligibility test is *not* definite and unequivocal.

Consider some examples, starting with (1a).

1a. If a thing is red (all over), then it isn't green (all over).

Since this seems to be a necessary truth, we ask, is its necessity determinable from logical form alone? At first glance, it would seem not to be, since the form of (1a) would seem to be something like (1b) (in standard notation), or (1c) (in tractarian notation); and we certainly can't determine truth from those forms.

1b. $\forall x\,(Rx \rightarrow \sim Gx)$
1c. N(x[N(N(Rx), N(Gx))])

But if we say that form alone doesn't determine that (1a) is necessary, then the intelligibility test will require us to say either that the sentence is nonsense or that it is used to make a contingent statement. Neither result seems correct.

Wittgenstein was aware of this problem, which he discusses at 6.3751. First look at the beginning and ending of the section.

> 6.3751 For two colors, e.g. to be at one place in the visual field is impossible, logically impossible, for it is excluded by the logical structure of color. Let us consider how this contradiction presents itself in physics. Somewhat as follows: that a particle cannot at the same time have two velocities; i.e. that at the same time it cannot be in two places; i.e. that particles that are in different places at the same time cannot be identical.
>
> (It is clear that the logical product of two elementary propositions can neither be a tautology nor a contradiction. The assertion that a point in the visual field has two different colors at the same time, is a contradiction.)

It seems evident that Wittgenstein neither classified (1a) as nonsense nor classified the statement it is used to make as contingent. Rather, he took the statement to be genuinely necessary, and the sentence to be meaningful in his extended sense. This requires him to deny that the statements *that o is red* and *that o is green* are elementary propositions, and also that (1b) or (1c) represent the real logical form of (1a). In effect, he conveniently invokes the doctrine of hidden logical form, and implicitly suggests that, at the level of logical form, the necessity of (1a) *is* a matter of its form alone.

This would be less worrisome if Wittgenstein had given a hint about what the real logical form of (1a) is. Perhaps the middle paragraph of 6.3751 provides the hint—namely, that the analysis of propositions about color is given by the physical theory of color. If so, the hint isn't helpful. The problem is to explain color incompatibility as logical impossibility. At most, the middle paragraph suggests that color incompatibility can be assimilated to physical impossibility—i.e., to the impossibility of (2a).

2a. o is at place p at time t and o is also at another place p' at time t.

But the apparent logical form of (2a) is just (2b), which is *not* formally contradictory.

2b. Lxpt & Lxp't

Thus, the problem of color incompatibility remains.[1]

This is just one example of a pervasive problem. As (2a) illustrates, our ordinary use of language is full of conceptual incompatibilities or necessities that are not in any obvious way determinable from the manifest linguistic form of the sentences used. To solve this problem, one would have to provide analyses in which the purely formal or structural properties of the logical forms of these sentences invariably revealed the conceptual incompatibilities and necessities holding among them. But Wittgenstein does not give such analyses, and provides few clues about how to come up with them.

In fact, the color incompatibility problem continued to trouble him for years. By 1929, he recognized, in "Some Remarks on Logical Form," that its solution required giving up the tractarian independence of elementary propositions, even though he continued to maintain much of the rest of the tractarian framework. At the time he wrote the *Tractatus*, he was so confident that his general principles must be correct that he thought that problems like color incompatibility must, somehow, be solvable. Since sentence (1a) is so obviously meaningful (in his extended sense) and since the statement it is used to make is so clearly necessary, he thought that it must have a logical form that shows it to be a tautology. Thus, the doctrine of the hiddenness of logical form was used to protect the intelligibility test from consequences deemed to be undesirable. This is, of course, a weakness of the test, since it leaves too much room for dispute about how to apply it.

Propositional attitude ascriptions like (3a) posed another problem.

3a. John believes (says/hopes/has proved) that the earth is round.

This sentence has another sentence 'the earth is round' as one of its constituents. According to the *Tractatus*, the only way for a meaningful sentence

[1] See pp. 90–91 of Fogelin (1987) for discussion.

R to occur in another meaningful sentence S is for S to be a truth function either of R by itself, or of R plus other sentences. The *Tractatus* maintains that all meaningful sentences are constructed by applying *truth-functional* operations to other sentences, and ultimately to the meaningful sentences (or uses of such sentences) that Wittgenstein calls *elementary propositions*.

> 5.54 *In the general propositional form propositions occur in other propositions only as bases of truth-operations.*

Since, in the *Tractatus*, the general propositional form tells us how all propositions are constructed, Wittgenstein is here claiming that the only way for a proposition p to have another proposition q as a constituent is for q to be one of the propositions to which truth-functional operators are applied in constructing p.

Examples like (3a) pose a threat to this doctrine. If (i) *sentence* (3a) is meaningful and (ii) the logical form of the statement it is used to make contains an occurrence of the sentence 'the earth is round', then the statement that the earth is round must be among the bases of the truth-functional operations used to construct the proposition that John believes/says/hopes/has proved that the earth is round. That could be so only if replacing 'the earth is round' in (3a) with any other true sentence would always preserve truth. Thus, according to the *Tractatus*, the proposition *that the earth is round* is a constituent of the statement (3a) is used to make only if the result of replacing 'the earth is round' with, say, 'arithmetic is reducible to set theory' is itself a sentence that is used to state a truth—i.e., only if the truth of (3a) (3b), and (3c) logically guarantees the truth of (3d).

3b. The earth is round.
3c. Arithmetic is reducible to set theory.
3d. John believes (says/ hopes/has proved) that arithmetic is reducible to set theory.

Since, in fact, the truth of (3d) is *not* logically guaranteed by the truth of (3a)–(3c), the doctrines of the *Tractatus* lead to the conclusion that either sentence (3a) is nonsense, or it is meaningful, but 'the earth is round' *doesn't* occur in the logical form of the statement (3a) is used to make. Since it is hard to envision what, in that case, the logical form of the statement might be, Wittgenstein is threatened with the result that the use of (3a) is meaningless and so its use fails to make any statement.

He addresses this problem at 5.541 and 5.542. In the immediately preceding section, he says, "*In the general propositional form propositions occur in a proposition only as bases of the truth-operations.*" He now adds:

> 5.541 *At first sight it looks as if it were also possible for one proposition to occur in another in a different way.*
> *Particularly with certain forms of proposition in psychology, such as 'A believes that p is the case' and 'A has the thought p', etc.*

For if these are considered superficially, it looks as if the proposition p stood in some kind of relation to an object A.
(And in modern theory of knowledge (Russell, Moore, etc.) these propositions have actually been construed in this way.)

5.542 *It is clear, however, that 'A believes that p', 'A has the thought p', and 'A says p' are of the form '"p" says p': and this does not involve a correlation of a fact with an object, but rather the correlation of facts by means of the correlation of their objects.*

In these passages, Wittgenstein claims that the real logical form of (3a) is different from what it first appears to be. Really, the logical form of examples of this sort is (4).

4. "p" says (that) p

Presumably, then, the logical form of (3a) is (5).

5. "the earth is round" says (that) the earth is round

Despite Wittgenstein's assurance that this *is clear*, his reasoning here is obscure. Nevertheless, we may be able to make something of it.[2]

He was probably thinking that when one believes something, one constructs a mental picture of a possible state of affairs—a representation of it. The representation is a fact, and the state of affairs represented is, as we may loosely put it, a possible fact. Since the one is a representation of the other, the elements in the facts are correlated with one another. In (3a), the expressions in the representing fact—i.e., the sentence 'the earth is round' or some mentalistic substitute for it—are correlated with things in the world that make up the nonlinguistic fact that the earth is round. That, in effect, is what (5) tells us.

Still, it's hard to credit Wittgenstein's explicit remark that (5) is the logical form of (3a). After all, (3a) specifies a specific agent, John, and a specific attitude, belief; (5) doesn't. There would be no change in (5) even if someone other than John were the agent, and the attitude reported were not belief but knowledge or assertion. Since (5) leaves out both the agent of (3a) and the particular attitude born to the postulated representation, (5) can't constitute the total content of (3a). But one might take (5) to be part of the logical form of (3a). To capture a belief attribution, one might understand (3a) as saying that John has formulated and accepted some representation that says that the earth is round. On this view the logical form of (3a) contains something along the lines of (5) as a part.

Although this seems to be interpretive progress, it doesn't help with our original problem. The sentence 'the earth is round' has an unquoted occurrence in (5) that is not one of the truth-functional bases of the statement

[2] See also Fogelin (1987), chapter 5, section 7.

(5) is used to make. If it were, we could replace that occurrence with an occurrence of any other true sentence, without changing truth value. But if we try this—e.g., by replacing the unquoted occurrence of 'the earth is round' in (5) with the sentence '2 + 2 = 4'—we end up with a falsehood.

6. 'The earth is round' says (that) 2 + 2 = 4.

Since substitution hasn't preserved truth value, we have the same trouble making the doctrines of the *Tractatus* compatible with the meaningfulness of (5), and the claim that it is used to make a genuine statement, as we had making them compatible with the meaningfulness of (3a), and the claim that it is used to make a legitimate statement.

So what was Wittgenstein's position? He seems to think that propositional attitude ascriptions like (3a) and semantic sentences like (5) are *not* really meaningful, and only appear to be used to make statements. On this view, a use of (5)—which was suggested as part of the analysis of (3a)—attempts to *state* something about the relationship between language and the world. But the relationship between language and the world *cannot,* according to the *Tractatus*, be meaningfully stated or described; it can only be shown.

Wittgenstein makes this point at 4.12–4.1212.

> 4.12 *Propositions can represent the whole of reality, but they cannot represent what they must have in common with reality in order to be able to represent it—logical form.*
>
> *In order to be able to represent logical form, we should have to be able to station ourselves with propositions somewhere outside logic, that is to say outside the world.*
>
> 4.121 *Propositions cannot represent logical form: it is mirrored in them. What finds its reflection in language, language cannot represent. What expresses itself in language, we cannot express by means of language. Propositions show the logical form of reality. They display it.*
>
> 4.1211 *Thus one proposition* 'fa' *shows that the object* a *occurs in its sense, two propositions* 'fa' *and* 'ga' *show that the same object is mentioned in both of them.*
>
> *If two propositions contradict one another, then their structure shows it; the same is true if one of them follows from the other. And so on.*
>
> 4.1212 *What can be* shown*, cannot be* said.

Here, Wittgenstein maintains that we can't use language to state or describe the relationship between language and the world that allows language to be meaningful, and that makes individual expressions mean what they do.

How shall we take this? Wittgenstein is right in thinking that there is no room for statements about the relationship between language and the world in the rigid system of the *Tractatus*, but he doesn't give an independent reason to think the view is plausible. Perhaps, it might be suggested on his behalf, to use and understand language you have to grasp the

relation between language and the world that makes your words meaningful; but once you have done so nothing about the relationship remains to be stated. But that isn't convincing. All that is established is that someone who didn't know any language couldn't *learn* language by being *told* what the relation between language and the world is. Such a person couldn't learn language that way because he couldn't understand the instructions. It is like saying you can't learn to read by reading a book that explains the reading process. There is nothing deep in this. Educational psychologists can discover the elements of reading and write them up for others to read. The same might be said for language in general.

For example, the sentence

7. 'Firenze' names Florence.

seems both meaningful and capable of being used to state a true proposition, even though its use says something about the relation between language and the world. Note, if I use the sentence

8. Bill is tall.

to tell you about a certain man's height, then I use the convention that the word 'Bill' names Bill to say something about him. Of course, my remark doesn't *state* the fact that the word 'Bill' names Bill. Rather, Wittgenstein would say that my use of (8) *shows* this. Okay, it does. He might add that no sentence, or use of a sentence, states *all* those facts about its own relation to the world that allow it to say what it does. Perhaps that is also correct. But it *doesn't* follow that no sentence can be used to state *any* of the facts about the relations between (a use of) its expressions and the world that allow it to say what it does. Nor does it follow that no sentence can be used to state a fact about the relationship between some expression and the world that allows *another* sentence to be used to say what it does. For example, there is no reason to deny that (9) is used to state a fact about the relationship between language and the world that is one of the facts that allows both (8) and (9) to be used to say what they do.

9. 'Bill' refers to Bill.

The lesson here is that although Wittgenstein's doctrines about what can't be expressed in language are overstated, they probably played a role in his seeming denial that sentences like (3a) (5), (7), and (9) are meaningful and capable of being used to make true statements. But there is a caveat to be added. In discussing propositional attitude reports illustrated by (3a), I assumed that, if they are used to make statements at all, then those statements contain constituent statements articulated by their complement clauses—e.g., by 'the earth is round' in the case of (3a). Although this assumption is extremely plausible, I don't see that it is dictated by the *Tractatus*. I don't see it is dictated, because I can't see that specific analyses of any sentences are dictated. We are told that all ordinary propositions

must be constructed by applying truth-functional operations to propositions about metaphysical simples, which, by design, are completely mysterious. This applies to ordinary propositions about everything we know anything about—people, houses, books, universities, automobiles, the earth, the sun, galaxies, black holes, sentences, propositions, attitudes to propositions, and so on. No such ordinary propositions are analyzed in the *Tractatus*, nor are we given the least clue about how to begin. So, if a dedicated tractarian wished to avoid the absurd conclusion that propositional attitude sentences are meaningless, and never used to make true statements about what is believed, asserted, and known, the dedicated acolyte might do so. This, I would say, is *not* a strength of the view.

2. THE LIMITS OF INTELLIGIBILITY: VALUE, THE MEANING OF LIFE, AND PHILOSOPHY

Consider the value statements, that happiness is good, that friendship is good, that causing pain unnecessarily is wrong, and that Michelangelo's *Pietà* is beautiful. Wittgenstein rejects the view that these are contingent, empirical propositions.

> 6.4 *All propositions are of equal value.*
>
> 6.41 *The sense of the world must lie outside the world. In the world everything is as it is, and everything happens as it does happen: in it no value exists—and if it did exist, it would have no value.*
>
> *If there is any value that does have value, it must lie outside the whole sphere of what happens and is the case. For all that happens and is the case is accidental.*
>
> *What makes it non-accidental cannot lie within the world, since if it did it would itself be accidental.*
>
> *It must lie outside the world.*

Wittgenstein doesn't give much by way of reason for rejecting the view that fundamental value judgments are contingent. But the rejection does seem plausible. Philosophers might disagree about the truth or falsity of many statements of value—statements that happiness alone is good, that taking an innocent life is always wrong, and that all other things being equal, lying is wrong—but it is hard to imagine these statements being true at some possible states of the world and false at others; it is also hard to imagine empirical observation and investigation being needed to find out whether the actual state of the universe is one that makes these statements true, or one that makes them false.[3] But if these value judgments are

[3] Of course, not all value statements are necessary and a priori, if true at all. For example, a claim to the effect that your speeding through a red light was justifiable, since your passenger was hemorrhaging, and would have died, had you not gotten her to the hospital when you did is clearly contingent and knowable only a posteriori, if true at all. But since propositions

neither contingent nor knowable only a posteriori, they also appear *not* to be tautologies (or contradictions). Value judgments are important to us and play a role guiding our actions that tautologies (and contradictions) don't. Moreover, if value judgments really were analyzable as tautologies (or contradictions), the truth (or falsity) of which were discoverable by their form alone, then presumably evaluative words like 'good', 'bad', 'right', and 'wrong' would have to be definable in terms of non-evaluative words. But by the time of the *Tractatus*, G. E. Moore had convinced most analytic philosophers that evaluative words were not definable.

According to the *Tractatus,* sentences containing evaluative words cannot be used to express genuine propositions. Thus, they are claimed to be senseless. If one person says "Murder is always wrong" and the other says "Murder is sometimes right," then neither has said anything true, and neither has said anything false. Wittgenstein's point is *not* that we can't find out which is correct, and which incorrect. His point is also *not* that no one can *prove* the correctness of his or her moral or other evaluative beliefs to a skeptic. His point is more radical: moral and evaluative sentences lack sense; they don't express propositions. Since there are no moral or evaluative propositions for us to believe, we don't have any moral or evaluative beliefs.

6.42 *So too it is impossible for there to be propositions of ethics.*

One can, of course, produce the words, "murder" "is" "always" "wrong," but one will *not* thereby have *said* anything more than if one had produced the words "procrastination" "drinks" "plentitude."

According to the *Tractatus*, there are no moral propositions; there are no moral beliefs, and there are no moral questions or problems. To think otherwise is to be confused about language. Once the workings of language have been laid bare, the traditional philosophical problems of value will not be solved; rather we will see that there never were any real problems there in the first place. From this a slogan was born: The philosophical analysis of language doesn't *solve* philosophical problems of value, it *dissolves* them.

It might seem that someone who characterizes all ethics and aesthetics as meaningless would regard ethical and aesthetic concerns as insignificant, and unworthy of serious attention. One imagines someone who thinks that what is important is giving an accurate scientific, or otherwise factual, description of the world. Since values don't fit into the description, they are unimportant. As we will see, that picture was associated with Carnap and other logical empiricists (though not entirely justly). But the picture was not, for good reason, associated with Wittgenstein. Although both

like this are not viewed by Wittgenstein as either elementary or truth functions of elementary propositions, they too are regarded as merely pseudo-propositions.

he and the logical empiricists thought of the realm of value as lacking in sense, Wittgenstein thought of it as very important non-sense. According to the *Tractatus*, all meaningful sentences are used to state tautologies, contradictions, or contingent truths or falsehoods that describe the way objects in the world are, or at least could be, combined. Although such sentences are meaningful, and are used to make statements that are true or false, Wittgenstein claimed not to find them very interesting or important. What was important and interesting, he thought, was how one lived one's life, what attitude one took toward things, and how one acted. According to the *Tractatus* these are matters about which it is impossible to say, or even to think, anything sensible.

> 6.423 *It is impossible to speak about the will in so far as it is the subject of ethical attributes.*
>
> *And the will as a phenomenon is of interest only to psychology.*
>
> 6.43 *If the good or bad exercise of the will does alter the world, it can alter only the limits of the world, not the facts—not what can be expressed by means of language.*
>
> *In short the effect must be that it becomes an altogether different world. It must, so to speak, wax and wane as a whole.*
>
> *The world of the happy man is a different one from that of the unhappy man.*

Wittgenstein is here being metaphorical, but one gets some idea of what he is saying. Consider the difference between the happy and the unhappy man. According to Wittgenstein, they might not differ in what they know or believe. Both might know all there is to know about science, history, psychology, or any other empirical discipline. They might believe the same things about inanimate objects, animals, other people, and even each other. Of course in certain cases they will express their beliefs differently. When the happy man believes that he is coming down with a cold, he will express this belief using the words "I am coming down with a cold," whereas the unhappy man will express that same belief about the happy man using the words "You are coming down with a cold." But, though their words are different, their beliefs are the same. Still, one man is happy and one is unhappy. The happy man wakes up in the morning with anticipation and a sense of well-being. He delights in his surroundings and his activities; treats other people kindly and considerately. The unhappy man feels and behaves in the opposite way. The difference between the two is, as Wittgenstein might say, at the level of value. It has nothing to do with what they think, or believe, or what they know to be true.

The suggested picture clashes with a venerable conception of philosophy. Philosophy has sometimes been thought of as a discipline that shares the highest aspirations of both science and religion. As highest science, its task has been thought to consist in the discovery of the most important and fundamental truths about reality, and the place of human beings in it. As deepest religion, its task has been taken to be the discovery of what

true excellence and happiness in human life consist in, and to tell us how to achieve them. These goals—describing reality and learning how to live the best life—have been thought by many to be not just compatible, but mutually reinforcing. The idea that excellence in the art of living is the result of knowing important truths about reality, oneself, and others, is an underlying presupposition of this view. Wittgenstein challenges this idea. The truth about how to live is *not* a deep and difficult mystery for the philosopher, or anyone else, to discover; nor is it a simple matter that we somehow know in advance. Excellence in living is not a matter of truth, knowledge, or belief at all. It is a matter of one's attitude, or response, to life. What attitude one adopts may be the most important thing in life, but it is not a matter of learning any facts.

It is hard *not* to be sympathetic with elements of this picture, which seem suggestive, insightful, and even true, however paradoxical that may sound to a strict tractarian. Much of the picture is distinctively Wittgensteinian, especially the seeming invitation to "something higher"—mysticism. But there is also an element that isn't unique to Wittgenstein, but rather was typical of the period in analytic philosophy in which he found himself, and through which he would live in the decades to come. The gulf between empirical fact and value that opened with Moore increased with Wittgenstein, grew larger with the logical empiricists, who were influenced by Wittgenstein, and continued in more sophisticated forms with later non-cognitivists. Philosophers during this period were *not* reluctant to make far-reaching *methodological* claims about ethics or other evaluative matters; they were *not* averse to pronouncing on what ethical or evaluative language was, or was not, all about. But they were very reluctant to argue *as philosophers* for substantive, controversial, or far-reaching normative theses of any kind, and they were anxious to sharply distinguish what they thought could be achieved in philosophy from anything of that sort.

Why this attitude was so widely shared during this period is an intriguing question for intellectual history and of historical sociology. Part of the answer is purely internal to the growing analytic tradition in philosophy—a matter of which philosophers, and which doctrines, were most compelling, and deservedly attracted the most attention. But part of the answer may have involved broader cultural currents—the rise of science, the decline of religion, the growth in wealth, the increase in urbanization, and the space for personal autonomy and freedom from traditional constraints thereby created. Whatever the ultimate causes, the absolute gulf between fact and value portrayed in the *Tractatus* was part of this current, including Wittgenstein's idiosyncratic take on it all.

To repeat, Wittgenstein adopts the paradoxical view that (i) if meaningful sentences are used to make genuine statements, which are either true or not, then what is expressed has nothing to do with value, and is not very significant to life, and (ii) if a sentence is used with the intention of stating something important about how we should live, then it will fail to

express anything that can even be thought. These views applied as much to religion, or to anything else connected to the meaning of life, as they did to ethical or other straightforwardly evaluative matters. Wittgenstein elaborates at 6.5 to 6.521.

6.5 *When the answer cannot be put into words, neither can the question be put into words.*
The riddle *does not exist.*
If a question can be framed at all, it is also possible *to answer it.*

6.51 *Skepticism is* not *irrefutable, but obviously nonsensical, when it tries to raise doubts where no questions can be asked.*
For doubt can exist only where a question exists, a question only where an answer exists, and an answer only where something can be said.

6.52 *We feel that even when all* possible *scientific questions have been answered, the problems of life remain completely untouched. Of course there are then no questions left, and this itself is the answer.*

6.521 *The solution of the problem of life is seen in the vanishing of the problem.*
(Is not this the reason why those who have found after a long period of doubt that the sense of life became clear to them have then been unable to say what constituted that sense?)

For Wittgenstein, ethics, religion, and talk about the meaning of life are relegated to the unsayable and unthinkable. About philosophy itself, the *Tractatus* is uncompromising. Just as the most fundamental ethical claims are neither tautologies nor contingent statements about empirically knowable facts, so philosophical claims are, in general, neither tautological nor contingent statements of empirical facts. Thus, like ethical sentences, they are nonsense. Hence, there are no meaningful philosophical sentences; there are no genuine philosophical questions; and there are no philosophical problems for philosophers to solve. It is not that philosophical problems are so difficult that we can never be sure we have discovered the truth about them. There is no such thing as the truth about them, because there are no philosophical problems.

What then is responsible for the persistence of the discipline of philosophy, and for the illusion that it is concerned with real problems for which answers might be found? The answer, according to Wittgenstein, is linguistic confusion. As he saw it, all the endless disputes in philosophy are due to confusion about how language works. If we could ever fully reveal the workings of language, these confusions would die out, and we would see the world correctly. When we did, we would see that there is no place in it for philosophy, just as there is no place for ethics. But this doesn't mean that there is nothing for philosophers to do. Philosophy can't properly aim at discovering true propositions; but, it can aim at clarifying the propositions we already have. Since Wittgenstein believed that everyday language disguises thought by concealing true logical form, he believed that the job of philosophy is to strip away the disguise and illuminate that form.

In articulating these views, the *Tractatus* was a key document in what was later called *the linguistic turn in philosophy*. Wittgenstein makes this clear at 4.11–4.112.

> 4.11 *The totality of true propositions is the whole of natural science (or the whole corpus of the natural sciences).*
>
> 4.111 *Philosophy is not one of the natural sciences.*
> *(The word 'philosophy' must mean something whose place is above or below the natural sciences, not beside them.)*
>
> 4.112 *Philosophy aims at the logical clarification of thoughts.*
> *Philosophy is not a body of doctrine but an activity.*
> *A philosophical work consists essentially of elucidations.*
> *Philosophy does not result in 'philosophical propositions', but rather in the clarification of propositions.*
> *Without philosophy thoughts are, as it were, cloudy and indistinct: its task is to make them clear and to give them sharp boundaries.*

According to the *Tractatus,* philosophy is linguistic analysis. Wittgenstein gives a clear statement of what he takes analysis to be in his first post-*Tractatus* paper.

> The idea is to express in an appropriate symbolism what in ordinary language leads to endless misunderstandings. That is to say, where ordinary language disguises logical structure, where it allows the formation of pseudo-propositions, where it uses one term in an infinity of different meanings, we must replace it by a symbolism which gives a clear picture of the logical structure, excludes pseudo-propositions, and uses its terms unambiguously.[4]

This conception of philosophy leads to the natural observation that, in the *Tractatus*, Wittgenstein didn't follow his own advice. He didn't produce a precise symbolism and use it to give analyses of sentences of ordinary language. He didn't do philosophy by producing the kind of analyses he says philosophers ought to produce. Rather, he practiced the kind of philosophy the *Tractatus* condemns as nonsensical. The *Tractatus* is filled with sentences that purport to make statements that are neither descriptions of contingent facts nor tautologies the truth of which is determined by their formal structure. Thus the *Tractatus* was nonsense by its own criteria.

This was not news to Wittgenstein.

> 6.53 *The correct method in philosophy would really be the following: to say nothing except what can be said, i.e. propositions of natural science—i.e. something that has nothing to do with philosophy—and then, whenever someone else wanted to say something metaphysical, to demonstrate to him that he had failed to give a meaning to certain signs in his propositions. Although it would not be satisfying to the other person—he would not have the feeling that we were teaching him philosophy—*this *method would be the only strictly correct one.*

[4] Wittgenstein (1929) at p. 163.

6.54 My propositions serve as elucidations in the following way: anyone who understands me eventually recognizes them as nonsensical, when he has used them—as steps—to climb up beyond them. (He must, so to speak, throw away the ladder after he has climbed up it.)

7. What we cannot speak about we must pass over in silence.

There are three ways of viewing Wittgenstein's final position. On one view, the *Tractatus* as a whole is self-defeating and/or self-contradictory, despite its illuminating insights on many points. Thus, the tractarian system must be rejected, and we should strive to find ways of preserving its insights while avoiding its inadequacies. This wasn't Wittgenstein's view. On the second view, the *Tractatus* is acceptable as it stands. In it, Wittgenstein deliberately violates the rules of language in an attempt both to *show* us what those rules really are and to reveal what the most basic knowledge of reality consists in. In order get us to *see* what the rules of intelligible thought and language are, he had to go beyond them. In order to make the most significant knowledge of reality available to us, he had to go beyond what can be stated to what he took to be a reality that can only be shown, not by any individual proposition, but by our entire system of propositions. Although I don't think this view is coherent, I do think it was Wittgenstein's view when he wrote the *Tractatus.* Finally, there is a third view, which has come into its own among some interpreters of Wittgenstein in recent decades. On this view, Wittgenstein deliberately set out to produce a compelling but clearly incoherent philosophical work, not to reveal any showable but unstateable truths, but to demonstrate, once and for all, the impossibility of philosophy.[5] I have been convinced by scholars who have closely examined not only the *Tractatus* itself, but also Wittgenstein's other writings, correspondence, and reported conversations, before and after the *Tractatus,* that this interpretation doesn't withstand scrutiny.[6]

Nor do I think it is charitable. The idea that an intentional descent into incoherence should have been expected to convince others that it was the upper limit of philosophical achievement strains credulity. Though some might take it as given that the young Wittgenstein's unique genius placed him at the summit of any past or future philosophy, I doubt that even his own legendary ego was quite that large. Nor do I understand how anyone who has worked through the many problems, difficulties, and misconceptions on display in our discussion in previous chapters of "the single great problem" of the *Tractatus*, the analysis of the proposition, could find such a view compelling. The most challenging difficulties are *not* meaningless doctrines that appear meaningful. The challenging problems are those in

[5] See Diamond (1988, 1991), Conant (1989, 1991, 2001, 2002).

[6] A powerful and, I believe, compelling critique of this interpretation of the *Tractatus* is provided in Hacker (2000). See also Proops (2001).

which understandable and sometimes promising ideas veer off into falsehood. All of this is lost in an interpretation that posits intentional but universal nonsense.

Such an interpretation obliterates the important advances—truths, or at least advances on the truth—contained in the *Tractatus*. The rejection of nonlinguistic Frege-Russell propositions, the embryonic conception of propositions as uses of sentences to represent things as standing in relations to one another, the embryonic theory of propositional truth as consisting in objects being the ways that true propositions represent them to be, the theory of truth-functional operators as operations rather than names of logical objects or constituents of facts, the "semantic" analysis of the tautologies of the propositional calculus (according to which they are all on a par) as opposed to axiomatic or other proof-theoretic accounts (according to which some truths of logic are more basic than others), and the attempt to extend this semantic account to logic as a whole, were all steps in the right direction. We read the *Tractatus* both to understand the historical impact of its insights, and to continue to learn from them today. All of this is obliterated in an interpretation that posits intentional and unmitigated self-refutation.

The correct view is, I believe, that the *Tractatus* is locally illuminating—both for its insights and its errors—despite being globally self-refuting. In addition to containing valuable insights, it is an object lesson in the absurdity of identifying five distinct types of truth—necessary truth, a priori truth, truth in virtue of meaning, logical truth, logically provable truth. Nothing was more significant in leading Wittgenstein down this disastrous path than his pre-Gödelian, pre-Tarskian conception of logic as the study not of *sentences* of formal languages, but of *propositions* expressed in both formal and natural languages. It is the latter, not the former, that are the objects of knowledge and necessity. It is the structurally simplest of the former, not the latter, that must be logically (but not necessarily or conceptually) independent, and that provide the basis for understanding and evaluating logically complex sentences.

It is a melancholy fact that the relationship between sentences and propositions is difficult, complex, and still insufficiently understood. But enough progress has been made to allow us to identify aspects of Wittgenstein's picture theory and his incipient analysis of propositions as *uses of sentences* as being the seminal breakthroughs they truly were. These breakthroughs were not wholly lost; they continued to play a role in Wittgenstein's later philosophy and in the "ordinary language" school of philosophy he helped to inspire. However, some of his tractarian insights were, until very recently, all but lost—and indeed eclipsed by the unfortunate tractarian identification of necessarily equivalent propositions—in the tradition in philosophical logic, formal semantics, and the philosophy of language leading from Frege and Russell to Carnap, Kripke, Montague, Lewis, Stalnaker, Kaplan, and others. Fortunately, that is no longer so.

Part Two

A NEW CONCEPTION OF PHILOSOPHY

LANGUAGE, LOGIC, AND SCIENCE

CHAPTER 5

The Roots of Logical Empiricism

1. The Origins of the Vienna Circle
2. Scientific Positivism: Comte and Mach
3. Developments in Logic, Mathematics, and Science: Hilbert, Poincaré, Duhem, and Einstein
4. Schlick's Early Epistemology and Philosophy of Science
5. The Kantian Legacy: Continuity and Reaction
6. The Impact of Wittgenstein

1. THE ORIGINS OF THE VIENNA CIRCLE

In August of 1929, a group of scientifically and mathematically inclined philosophers identified itself as *The Vienna Circle* in a proclamation dedicated to Moritz Schlick entitled "The Scientific Conception of the World." The proclamation was written under the auspices of the Ernst Mach Society by three of its members, Hans Hahn, Otto Neurath, and Rudolf Carnap. Announcing what it took to be a new, scientifically based conception of philosophy, it ended with a list of the members of the Vienna Circle—which included Gustav Bergmann, Rudolf Carnap, Herbert Feigl, Philipp Frank, Kurt Gödel, Hans Hahn, Viktor Kraft, Otto Neurath, Moritz Schlick, Friedrich Waismann, and four others. There was also a *list of those sympathetic to the Vienna Circle*—which included Kurt Grelling, F. P. Ramsey, Hans Reichenbach, and seven others—plus a list labeled *leading representatives of the scientific world-conception*—Albert Einstein, Bertrand Russell, and Ludwig Wittgenstein. Although the Circle was never philosophically homogeneous, it gave birth to a highly influential school of philosophy known as *logical positivism* or *logical empiricism*, before its geographical dispersal less than a decade after the formal announcement of its existence.

Though the circle announced its program to the world in 1929, its origins can be traced to Ernst Mach's tenure as the first holder of the Chair in Philosophy of the Inductive Sciences at the University of Vienna, 1895–1901.

His tenure was followed in the first decade of the twentieth century by regular meetings in Vienna of scientific thinkers—led by Philipp Frank, Otto Neurath, and Hans Hahn—who had been influenced by Mach, Duhem, and Poincaré. When, in 1922, Moritz Schlick was brought by Hahn to Vienna to occupy Mach's old chair, what was to become the Vienna Circle found a vigorous and charismatic leader. Schlick was a leading epistemologist whose early work sought both to interpret Einsteinian physics and to draw far-reaching lessons from it about the nature of human knowledge. This work combined strains of verificationism and scientific realism existing in an uneasy tension with one another. However, this was only the beginning for Schlick, who was eventually to evolve into an ardent verificationist—in part due to the influence of Carnap's 1928 *Logical Structure (Aufbau) of the World* (written between 1922 and 1925), in part due to Schlick's, and other circle members', reading of the *Tractatus*, and in part due to direct interaction with Wittgenstein beginning in 1927 and continuing until the mid-1930s.

2. SCIENTIFIC POSITIVISM: COMTE AND MACH

The term "positivism" names an intellectual tradition emphasizing the practical nature of science and its importance in human life. It dates back to the French philosopher Auguste Comte (1798–1857). Comte's *Cours de philosophie positive* (1830–42), translated as *Positive Philosophy*, traces the history of human thought as progressing through three stages—the theological, the metaphysical, and the positive (scientific) stages. The goal of the first two was to attain knowledge of first and final causes of "phenomena" by postulating either agents or forces. Comte explains the positive stage as one in which

> In the final, positive state, the mind has given over the vain search for Absolute notions, the origin and destination of the universe, and the causes of phenomena, and applies itself to the study of their laws – that is, their invariable relations of succession and resemblance.[1]

This shift in subject matter—from the unknown and putatively unknowable to the humanly discoverable—is characteristic of positivism. For the positivist, the goal of science is to identify the most encompassing true generalizations about "phenomena" under investigation, as opposed to unearthing hidden, but metaphysically real, causes. Comte's other major idea was that science should be thought of as a single unified inquiry. For him, divisions between different sciences were largely superficial. Although individual sciences may deal with different classes of phenomena, he took

[1] Comte (1830–42), vol. 1, p. 2.

their aims and methods to be essentially the same—to discover regularities by observation, hypothesis formation, and test. Not having a set of phenomena of its own to study, abstract mathematics was seen not so much as a special science, but as an essential tool of all sciences. Geometry was the exception for Comte, who viewed it as the abstract study of physical space.

Like later positivists, his conception of science extended to biology, psychology, political science, and sociology—including all aspects of the study of human beings and human society. Sometimes called "the father of sociology," he not only believed in a science of human society, but also in its preeminence among the sciences—not in being the most advanced, but in being the most encompassing and important. He thought it the most encompassing because its results rested upon those of other sciences and also because science itself, being an institutionalized form of inquiry aimed at furthering the common good, fell within the domain of sociological study, and in principle could properly be regulated by it. Comte went so far as to task sociology with instituting a *religion of humanity*, in which God would be replaced by humans-in-society as an object of reverence, and in which institutional forms of a religious character would promote love of, and service to, humanity. Though later positivists didn't follow him in this, many shared his animus toward traditional religion, his zeal in promoting an all-encompassing scientific worldview, and his understanding of science itself, no matter how abstract or abstruse, as posing solvable problems the solutions to which would advance human well-being.

The most important figure connecting the later advocates of *logical* positivism to this earlier version of positivism was Ernst Mach (1838–1916). A distinguished physicist, historian, and philosopher of science, his wide-ranging interests led him to deep involvements with evolutionary biology, psychology, and psychophysiology. While his early criticism of Newton's conception of absolute space and time won praise from Albert Einstein and Max Planck, his implicit verificationism and resolute anti-realism about unobservable entities, illustrated in his anti-atomism and initial opposition to the kinetic theory of heat, earned him harsh criticism from Planck. Like Comte, he was committed to the unity of science, which he viewed as an instrument of human advancement. Unlike Comte, he thought of this advancement as a step in the evolution of humanity's biological adaptability.

> The biological task of science is to provide the fully developed human individual with as perfect a means of orientating himself as possible. No other scientific ideal can be realized, and any other must be meaningless.[2]

Just as Comte conceived of science as studying regularities among "phenomena," so Mach thought of it as studying what he called "sensation."

[2] Mach (1914), p. 37.

These were cognitive events or products resulting, in the case of vision, from the effect of light on the retina. Although this view may sound physicalistic, with cognitions conceived as neural events, physicalism was not the whole story, since the retina itself was, for Mach, simply a complex of sensations. Indeed all of science, including psychophysiology, was about these cognitive events or products. According to Mach, sensations are the simplest constituents of sense experience—visually experienced color, shape, size, and distance, tactilely experienced shape, size, and texture, auditorily experienced sound, motor sensations of effort and force, plus pains, pleasures, and emotions. The properties of these elements always depend at least in part on the experiencer. The elements themselves are intrinsically neither mental nor physical. Rather, they are assigned to these categories only in inquiries that relate them to one another either (i) as constitutive parts of a single stream of consciousness, in which case they are called "sensations" and regarded as psychological, or (ii) as constitutive parts of complexes not all the elements of which need belong to a single stream of consciousness, in which case they are regarded as physical.

> The great gulf between physical and psychological research persists only when we acquiesce in our habitual stereotyped conceptions. A color is a physical object as soon as we consider its dependence, for instance, upon its luminous source, upon other colors. . . . When we consider, however, its dependence upon the retina . . . it is a psychological object, a sensation. Not the subject matter, but the direction of our investigation, is different in the two domains.[3]

In short, Mach was neither an idealist nor a physicalist, but a neutral monist in the sense explored by Russell in the final chapter of *The Philosophy of Logical Atomism*.[4] His fundamental elements, out of which reality is entirely constructed, are, like those of the Berkeleyan idealist, conscious cognitive events or products. Unlike the idealist, however, he takes these elements to be the building blocks out of which not only the physical world, but also "the self," are constructed. Because Machian elements are experiences that are conceptually prior to the experiencing subject, they are not modifications of an antecedent consciousness, but free-floating cognitions, of which the subject is merely a collection or construction. Psychology studies this construction; physical science studies the construction of physical things out of the very same elements; psychophysiology studies the connection between mind and body. For Mach, who contributed to each of these disciplines, this was the ultimate *unity of science*.

With an unintended irony not unknown in philosophy, Mach combines his highly revisionary metaphysics, founded on an a priori conception of

[3] Ibid., pp. 18–19.
[4] See Soames (2014), chapter 12, section 9, particularly pp. 625–29, where Russell's exploration of how to transform his phenomenalism into a system of neutral monism is discussed.

experience and observation, with a vigorous rejection of all a priori metaphysics. Thus, in the preface to the second edition of *The Analysis of Sensations* we are told:

> One and the same view underlies both my epistemologico-physical writings and the present attempt to deal with the physiology of the senses—the view namely that all metaphysical elements are to be eliminated as superfluous and as destructive of the economy of science.[5]

and in the preface to the fourth edition:

> The opinion . . . that science ought to be confined to the compendious representation of the actual, necessarily involves as a consequence the elimination of all superfluous assumptions which cannot be controlled by experience, and, above all, of all assumptions that are metaphysical in Kant's sense.[6]

It is instructive to compare the scientific anti-realism in these passages with a similar anti-realism that arose decades later in Russell's version of phenomenalism. The motivations for the two phenomenalist systems were different. Mach wished to obliterate any substantive distinction between the mental and the physical; Russell wished to repudiate skepticism by explaining empirical knowledge. But their broadly anti-realist visions were similar. Both began by eliminating the supposedly superfluous metaphysical *element of hypothesis* in our conception of ordinary material objects as substances that persist through time and changes in their observable properties, and exist independently whether or not they are perceived.[7] This accomplished, they characterized ordinary observable objects as *constructions* (or, for Russell, *logical constructions*) out of sensations. This cleared the way for treating all unobserved entities in science as themselves constructions.[8]

Here is a sample of Mach's thoughts on the subject.

> We must regard it as an additional gain that the physicist is now no longer overawed by the traditional intellectual implements of physics. If ordinary "matter" must be regarded as a highly natural, unconsciously constructed mental symbol for a relatively stable complex of sensational elements, much more must this be the case with the artificial hypothetical atoms and molecules of physics and chemistry. The value of these implements for their special, limited purposes is not one whit destroyed. As before, they remain economical ways of symbolizing experience.[9]

[5] Mach (1914), p. x.
[6] Ibid., p. xii.
[7] See chapter 1 of Mach's (1914) *The Analysis of Sensations* and chapter 3 of Russell's (1914b) *Our Knowledge of the External World* (discussed in Soames [2014], chapter 11).
[8] Compare Mach (1914), chapter 14, with Russell (1914b), chapter 4.
[9] Mach (1914), p. 311.

> Now one might be of the opinion, say, with respect to physics, that the portrayal of the sense-given facts is of less importance than the atoms, forces, and laws which form, so to speak, the nucleus of the sense-given facts. But unbiased reflection discloses that every practical and intellectual need is satisfied the moment our thoughts have acquired the power to represent the facts of the senses completely. Such representation, consequently, is the end and aim of physics, while atoms, forces, and laws are merely means facilitating the representation.[10]

> The cogency of geometry (and of all mathematics) is due, not to the fact that its theories are arrived at by some peculiar kind of knowledge, but only to the fact that its empirical material which is particularly convenient and handy, has been put to the test very often, and can be put to the test again at any moment.[11]

Though the logical positivists who were to succeed Mach differed from him in many ways, most of his fundamental themes were eventually to become theirs as well, including the unity of science, the centrality of observation, the desire to overcome psychophysical dualism, the temptation of phenomenalism, a tendency toward verificationist anti-realism, the rejection of absolute space and time, and the rejection of geometry as the abstract, a priori study of physical space.[12]

3. DEVELOPMENTS IN LOGIC, MATHEMATICS, AND SCIENCE: HILBERT, POINCARÉ, DUHEM, AND EINSTEIN

When Mach wrote *The Analysis of Sensations*, he did not have the benefit of Frege's new logic or the elegant Frege-Russell logicist vision of mathematics. Since it is hard to overestimate the philosophical importance of the new logic, including its role in the paradigm case of a philosophically significant reduction of one theory to another, it marks what is, perhaps, the most telling difference between traditional scientific positivism and the rising school of logical positivism, or as it came more frequently to be called, logical empiricism. Certainly Rudolf Carnap's grand scheme of unifying science by systematically reducing theories of one scientific domain to another (examined in chapter 6) could not have been pursued without it. When another significant development was added—Wittgenstein's identification of apriority and necessity with analyticity—putatively meaningful discourse about the world falling outside the purview of any

[10] Ibid., pp. 314–15.
[11] Ibid., p. 346.
[12] One of Mach's key themes that was regrettably not much taken up by the logical positivists involved new developments of biology— from Darwin's evolutionary perspective to the scientific study of genetics.

possible unification of science came to be viewed as not merely idle but meaningless.

For the early logical positivists—particularly Schlick, Reichenbach, and Carnap—geometry provided a bridge between logic and mathematics, on the one hand, and physics, on the other. By the turn of the twentieth century, two familiar Kantian ideas were under attack—that geometry must be Euclidean and that it is the a priori study of a fundamental aspect of the empirical world. As noted in volume 1, non-Euclidean geometries had been around for decades, prompting speculation that physical space may itself be non-Euclidean. Nevertheless, Frege remained a Kantian about it, exempting geometry from his logicist reduction and continuing to regard Euclidean geometry as the synthetic–a priori truth about experienced space. However, many others—including Mach, Hilbert, Poincaré, Duhem, and Schlick—did not follow suit.[13]

In 1899, David Hilbert's *Foundations of Geometry* (*Grundlagen der Geometrie*) demonstrated that formal reasoning in axiomatized geometric theories need not appeal to any intuitive conception of space.[14] Thought of in this way, geometry is purely abstract and mathematical, whether Euclidean or not, and so has no intrinsic relation either to intuitively experienced space or to physical space. Henri Poincaré agreed, while noting that when a geometry is incorporated into an empirical theory, its role is not to represent any aspect of physical reality, but to facilitate the generation of correct empirical predications.[15] For him, the geometry of a theory was, in effect, an elaborate convention for getting from one empirical data point to another. Since in any given case there may be alternative conventions the adoption of which would yield equivalent empirical results, none is uniquely required in order to achieve objective scientific truth. In such a case, the proper choice among empirically equivalent alternatives is the one that achieves the greatest theoretical simplification.

In characterizing hypotheses about the unobservable as matters of *convention*, as opposed to matters of *fact*, Poincaré likened them to stipulative definitions—leading him to think, for example, that the Newtonian law that the acceleration of a freely falling body is constant can never be empirically falsified because it simply defines the concept *freely falling body*.[16] (If we find that the acceleration of a falling body is not constant, we don't reject the hypothesis, but look for some previously unnoticed force that must be acting on it.) Another of his contemporaries, Pierre Duhem, who shared Poincaré's positivistic conception of scientific theories, conceived of the failure of an individual scientific hypothesis to state directly testable facts rather differently. According to Duhem, such a failure doesn't show

[13] See Soames (2014), pp. 43–44.
[14] Hilbert (1899).
[15] Poincaré (1902).
[16] See Alexander (1967), p. 362.

that the hypothesis is really an analytic definition, but rather illustrates the general point that non-observational statements of a theory are *never* individually falsifiable, because they *always* require subsidiary hypotheses to generate observational predictions.[17] For Duhem, it was theories, not individual hypotheses, that may be confirmed or disconfirmed by observational evidence. But the larger point survives. Like Poincaré, he both divorced geometrical theories from Kantian spatial intuition and thought of them as interpretable only via embedding in an encompassing physical theory. Hence, they could no longer be presumed to be either Euclidean or instances of the synthetic a priori.

Mach, Poincaré, and Duhem are examples of scientist-philosophers who influenced the school of logical empiricism that grew out of the Vienna Circle. But the most powerful scientific influence was provided by Albert Einstein, whose theories of special and general relativity relativized the Newtonian notions of space, time, and mass, while also according physical reality to the non-Euclidean geometries of certain physical systems. One can get some sense of the change by considering how the temporal simultaneity of two events occurring at a distance from one another is established. In daily life we judge two nearby events in our visual field to be simultaneous when we see them at the same time—when light emanating from one impacts our eyes at the same time as light emanating from the other. Since the distances are typically so short in relation to the speed of light, this method works well for everyday purposes. But when we let the distances of the events from each other, and from the observer, vary, and get arbitrarily great, we need a method for determining the time it takes each ray of light to reach our eyes. Section 1 of Einstein's 1905 paper "On the Electrodynamics of Moving Bodies" deals with this problem in a single inertial frame (where we don't have to consider the motions of any objects other than those within a limited physical system).[18] The paper, which introduces the special theory of relativity, modifies our understanding of the relation of temporal simultaneity.

The central idea can be vividly illustrated by imagining synchronized clocks present at the sites of two events A and B located at arbitrary distances from each other and from an observer. Each clock starts the moment when its paired event occurs. The clocks are then transported to the observer through different spatial paths at different speeds. If the speed of their transmission through space didn't affect their running, then an observer who knew how far they traveled could simply check their readings when they arrived. If one went twice as far but moved twice as fast, the events would be simultaneous if the clocks registered the same time when they riched the observer. According to relativity theory, however,

[17] Duhem (1914).
[18] Einstein (1905).

the clocks' behavior *is* affected by the speed of their transmission through space.[19] If this sounds incoherent, it is probably because one is thinking of clocks as metaphysical know-not-what's that, *by definition,* track the passage of time, which, *by definition,* exists independently of any physical phenomenon. But that thought is unfounded. It's not true a priori that there *must* be such a thing as time *conceived of in that way.* The clocks imagined in the example are physical mechanisms, and so are subject to physical laws. Because of this, it's not true a priori that their behavior will be unaffected by the speed they move through space. Relativity theory maintains that their behavior *is* affected, thereby threatening the pretheoretic notion of simultaneously occurring events at a distance.

It is instructive to examine what happens when we replace (or sharpen) this pretheoretic idea with a physically defined notion of the simultaneity applying to events at a distance. Let us say that for events at a distance to be *physically simultaneous*, and so *not* separated in time, is for there to be *no possible causal connection* (e.g., by light from one reaching the other) between them. The argument of Einstein's 1905 paper shows that although physical simultaneity, so understood, is a symmetric relation, it is *not* transitive. This result is illustrated by a certain sequence of events—A, B, C, and D—*all occurring in that temporal order at point 1 in space* and an event Δ occurring at some spatially distant point 2. In the example, a ray of light travels from A to Δ, with Δ later than A, and a ray of light travels from Δ to D, with Δ earlier than D. Because the transmission of light is not instantaneous, events B and C, which occur at point 1 after A but before D, can't be connected by rays of light to the occurrence of Δ at point 2. (Since B follows A, light from B can reach point 2 only after Δ has occurred, and since C precedes D, light from Δ can't reach point 1 at the moment prior to D at which C occurs.) The basic relativistic result is that there are *no physical relations of any kind* capable of causally connecting event Δ at point 2 with *any events occurring at point 1 after A and before D*.[20] This means that events B and C at point 1, which occur after A but before D, are both *physically simultaneous* with Δ at point 2, even though B temporally precedes C.

Since we don't want one event to be *simultaneous* with two temporally nonoverlapping events one of which is later than the other, we need to adjust our understanding of these relations. One way to do so is let the relations *simultaneous with, before,* and *after* to be undefined for pairs one of which is Δ and the other of which is any event in the temporal interval from A to D at point 1. If we do, then these temporal relations will be physically grounded, but only partially defined. A different way out is to choose a unique event in the range of indeterminacy at point 1 and

[19] The example is nicely explained on p. 134 of Grunbaum (1967).
[20] Ibid., pp. 134–35.

simply stipulate that it is to count as *the event at point 1* that is simultaneous with Δ at point 2 (within a single inertial frame). The adoption of such a rule means that the simultaneity relation embedded in the theory will be partially conventional, rather than one the obtaining of which between arbitrary events is entirely determined by objective physical facts.[21] This seeming disadvantage is offset by the fact that when one considers not a single inertial system but all points in all inertial systems, the simplicity achieved by having a single uniform rule is significant. For this reason, Einstein offered a conventional synchronization rule for simultaneity at a distance (within the range of actual physical indeterminacy) for all relevant pairs of events at a distance in all inertial systems. The convention is that Δ is simultaneous with the midpoint of the segment A–D, as measured by an ideal clock carried by an inertial observer who passes through both points. This makes simultaneity relative to a reference frame, because the choice of the midpoint of A–D depends on the fact that A and D are at the same point in space, which is relative to the reference frame. In this way, Einstein's convention makes simultaneity relative to a reference frame. This allows the speed of light to be held invariant at 186,000 miles per second across all systems, though different values could have been assigned had different conventional choices been stipulated.[22]

When one considers different inertial systems S1 and S2 moving with respect to one another, the spatial extension of a rigid rod will depend on the criterion of simultaneity in the systems and the position and velocity of the motion of one system relative to the other. As a result, the extension of the rod becomes relativized to the reference frame (inertial system) in which the question is considered. A rod of length L at rest in S2 that is aligned with the direction of velocity of system S2 with respect to system S1 will, in S1, be a moving rod with a length less than l, where the length of the moving rod in S1 is determined by certain *simultaneous* events involving its end points. This illustrates the relativization of Newtonian notions like spatial extension and mass in the new physics.[23] The explicit attribution of non-Euclidian geometries to specific physical systems—in particular to rotating discs—is found in Einstein's 1916 paper on the general theory of relativity.[24] This doesn't mean that space in general is non-Euclidian. It means that there is no geometry, Euclidian or non-Euclidian, that is determined independently of the distribution of matter in particular physical systems and the physical relationships of these systems to one another.

[21] Ibid., p. 136.
[22] Ibid., p. 138. Thanks to an anonymous referee for helping me clarify the material in this paragraph.
[23] Ibid., pp. 138–39.
[24] Einstein (1916).

4. SCHLICK'S EARLY EPISTEMOLOGY AND PHILOSOPHY OF SCIENCE

The founding figure of the Vienna Circle, Moritz Schlick, studied physics under Max Planck at the University of Berlin, and received his Ph.D. in 1904. After a year of experimental work in physics, he turned to the study of philosophy, holding positions at Rostock and Kiel from 1910 to 1922, when he moved to Vienna to occupy Mach's old chair. The chief philosophical influences on him during this period were Einstein's theories of special and general relativity, which were central topics in his most important early work, including "The Philosophical Significance of the Principle of Relativity" (1915), *Space and Time in Contemporary Physics* (1917 [1979]), and *General Theory of Knowledge* (1918 [1985]). Schlick's perspective in these works is illustrated by the following two passages from *Space and Time in Contemporary Physics* in which he contrasts both Newtonian physics and special relativity (described in the first passage) with general relativity (described in the second).

> [Space] still preserved a certain objectivity, so long as it was still tacitly thought as equipped with completely determined metrical properties. In the older physics one based every measurement procedure . . . on the idea of a rigid rod, which possessed the same length at all times, no matter at which place and in which situation and environment it may be found, and, on the basis of this thought, all measurements were determined in accordance with the precepts of Euclidean geometry. . . . In this way, [the structure of] space . . . was thought to be entirely independent of the physical conditions prevailing in space, e.g., . . . of the distribution of bodies and their gravitational fields.[25]

> If we want . . . to maintain the general postulate of relativity in physics, we must refrain from describing measurements and situational relations in the physical world with the help of Euclidean methods. However, it is not that, in place of Euclidean geometry, a determinate other geometry . . . would now have to be used for the whole of space, so that our space would be treated as pseudo spherical or spherical. . . . Rather the most various kinds of metrical determinations are to be employed, in general, different ones at each position, and what they are now depends on the gravitational field at each place.[26]

Schlick understood the new physics and embraced the independence of its fundamental spatial and temporal concepts from our ordinary ones, whether *intuitive* in the Kantian sense or simply pretheoretic. "Intuitions," in the post-Kantian continental philosophy of Schlick's milieu, referred to raw, conceptually unstructured sensory inputs, which, in the Kantian

[25] Schlick (1917 [1979]), pp. 238–39.
[26] Ibid., at p. 240.

picture, are structured by the "pure forms" of spatial and temporal "intuition." Thought by Kant to be Euclidean, the "intuitive" space of our visual (and conceptually imaginable) experience can't be the physical space of general relativity—both because Einsteinian space is more abstract than either Euclidean or non-Euclidean space and because the spatial concepts that occur in relativity theory are not "intuitive" concepts at all. Instead, they are concepts the contents of which are holistically determined by their role in a complex and broadly encompassing physical theory. According to Schlick, we don't grasp these physical concepts by *first* grasping "intuitive" concepts that apply to the deliverances of the senses—whether spatial, temporal, or qualitative—and then defining the physical concepts in terms of the intuitive (perceptual) ones, as Russell attempted to do in *Our Knowledge of the External World*.[27] Rather, our grasp of the physical concepts is supposed to coincide with our understanding of the total theory in which they play significant parts.

What, one wonders, is such an understanding supposed to amount to? In chapter 1 of *The General Theory of Knowledge,* Schlick takes Hilbert's purely formal, axiomatic treatment of geometry as the model for conceptualizing scientific knowledge. Michael Friedman explains Hilbert's significance for Schlick as follows.

> Just as the Hilbertian focus on formal-logical structure is intended to purge geometrical deduction from possibly misleading reliance on spatial intuition, so as . . . to allow the logical relations of dependence between geometrical propositions to stand out more clearly, Schlick's theory of scientific conceptualization is intended to free it . . . from all vagaries of intuitive representation by allowing us to characterize scientific concepts . . . solely in terms of their formal-logical relations to one another. In this way, the distinction between a formal axiom system for geometry (what we would now call an uninterpreted formal system) . . . and a possible interpretation for such a system via intuitive spatial forms . . . provides Schlick with the primary model for his own distinction between knowledge and experience, or acquaintance.[28]

In one way this explanation is illuminating, but in another way it is (as Friedman realizes) puzzling. Just as Hilbert's formalization helped liberate geometry from unfounded aprioristic assumptions about the nature of its subject matter and application, so the highly abstract laws of general relativity theory helped liberate physics from unfounded aprioristic assumptions about the physical world, which is its subject matter. But whatever one thinks of mathematics in general, and geometry in particular, our knowledge of *physical theory* is *not* knowledge of an uninterpreted formal system. Nor is the "interpretation" of physical theory, which gives rise to

[27] See Soames (2014), pp. 545–554.
[28] Pages 34–35 of the reprinting of Friedman (1983) in Friedman (1999).

our knowledge of its subject matter, anything like what we now regard as the (model-theoretic) interpretation of a purely formal system. The latter consists of (i) an independent specification of a domain of objects, (ii) a mapping of the vocabulary of the system onto various objects of, and set-theoretic constructions out of, the domain, and (iii) a recursive account of the truth conditions of the sentences of the formal system, the specification of which requires *antecedently grasped concepts* to interpret the vocabulary of the system. Whatever "interpreting" a physical theory may be, it is not like that. Thus, we urgently need to know what its interpretation does consist in, and how, for Schlick, our knowledge of the "interpreted" theory is supposed to provide the general model for all knowledge.

The key idea, suggested in section 7 of *The General Theory of Knowledge*, is that theory interpretation, and the knowledge we derive from it, is the result of *implicit definition*, which involves taking theoretical primitives to mean whatever they have to mean in order for the axioms of the theory to be true. Here is Schlick.

> The meaning and effect of implicit definitions and how they differ from ordinary definitions ought now be more clear. In the case of ordinary definitions, the defining process terminates when the ultimate indefinable concepts are in some way exhibited in intuition (concrete definition). This involves pointing to something real, something that has individual existence. Thus we explain the concept of "point" by indicating a grain of sand, the concept of "straight line" by a taut string, that of "fairness" by pointing to certain feelings that the person instructed finds pleasant in his own consciousness. *In short, it is through concrete definitions that we set up the connection between concepts and reality. Concrete definitions exhibit in intuitive or experienced reality that which henceforth is to be designated by a concept*. On the other hand, implicit definitions have no association or connection with reality at all; specifically and in principle they reject such association; they remain in the domain of concepts. A system of truths created with the aid of implicit definitions does not at any point rest on the ground of reality. On the contrary, it floats freely, so to speak, and like the solar system bears within itself the guarantee of its own stability. *None of the concepts that occur in the theory designate anything real; rather they designate one another in such fashion that the meaning of one concept consists in a particular constellation of a number of the remaining concepts*.[29]

This passage is a mixed bag. On the positive side, we learn that the primitive concepts of interpreted physical theory are not "intuitive" concepts applying to private conscious experiences arising from the deliverances of our senses, nor, it would seem, are they pretheoretic concepts of any sort that we grasp independently of the theory. Rather, they are the constituents of a self-contained network of interdependent concepts ("like the solar system")

[29] P. 37 of Schlick (1918 [1985]), my emphasis.

the relationships between which are explicated by the theory itself. On the negative side, Schlick's appeal to implicit definition is, at best, unartfully expressed, and at worst, absurd. The concepts expressed in physical theory don't "designate one another," and, if it is true that they "don't designate anything real," we are owed an explanation of what this amounts to.

The most natural explanation starts from the observation that implicit definition of theoretical primitives requires a partial interpretation of the theory already to be in place. If every nonlogical constant in the theory were up for interpretation, the stipulation "Let theoretical primitives mean whatever they must in order for the theory to be true" could never yield determinate content. When the theory is empirical it is all but irresistible to take observational predicates (and other vocabulary) appearing in the observational statements entailed by the theory to be antecedently understood. Doing so provides us with a basis for interpretation. If we then stipulate *Let the non-observational primitives mean whatever they must mean in order for the observational consequences of the theory to be true*, we can begin to inquire what the truth of a theory consists in.

Is it enough that all observational consequences be true? What about theories that entail one or more observational falsehoods? Such a theory will be false, but surely not meaningless or uninterpreted. Even if all the observational consequences of a theory are true, might there be *different* meaning assignments to the non-observational primitives that would make all statements of the theory true? What should be said about alternative theories—with radically different, even inconsistent, non-observational statements—that make precisely the same observational predications? Are we to take one of them to be true and the others false, if their observational contents coincide and their non-observational vocabularies are initially uninterpreted?

Taking these quandaries together, one is tempted to identify a theory's meaning or content with the content of its observational predictions, thereby embracing holistic verificationism. Although this would fit Schlick's emphasis on epistemological holism, while (perhaps) vindicating his startling comment that "none of the [non-observational] concepts that occur in the theory designate anything real," it would not do justice to the evident tension between verificationism and scientific realism in *The General Theory of Knowledge*. Nor would it accommodate his surprising doctrine that the contents of our sense experiences—which he takes to provide the basis for our construction of physical content—are themselves *objectively unknowable* and *incommunicable* until they are subsumed under the physical concepts the construction of which they (supposedly) make possible.

Schlick discusses the relationship between private sensory content and objective physical content—which he calls *transcendent*—in section 31.

> The ordering in space and time of the contents of consciousness is . . . the means by which we learn to determine the transcendent ordering of the things

> that lie beyond consciousness. This transcendent ordering is the most important step toward a knowledge of these things. . . . We saw that establishing an identity—this is what all knowledge consists in—means, as far as external things are concerned, locating things at the same point in time and space. Everything in the external world . . . is at a particular place at a particular time; and to find one thing in another is ultimately to assign to both of them the same place at the same time. We must now make this definition more precise by specifying that when we use the expressions 'space' and 'time', we mean the transcendent ordering of things. . . . The important thing now is to get clear about how we proceed from the intuitive spatio-temporal ordering to the construction of the transcendent ordering.[30]

Put aside any notion that the ordering being "constructed" is the product of our minds; it isn't. The task is to explain how our *intuitive* (Kantian) *conceptions* of spatial and temporal relations in private experience allow us grasp of the concept of objective space-time points. *Since these points are the ultimate subject matter of Einsteinian physics, this "construction" of the physical order from the subjective order provides the interpretation of physical theory.* Though Schlick's discussion is neither complete nor precise, his idea is, I think, that the primitives of physical theory are to mean whatever they must mean in order for the claims about objective space-time points made by the theory to be true. Since it is not required that each of these claims must, in principle, be observationally verifiable by us, there is an element of scientific realism in his philosophical outlook in this early period.[31]

Schlick calls his method of "constructing" the physically objective out of the intuitively subjective the *method of coincidences*, which he models on physical measurements in which the length of an object is determined by correlating points on a measuring rod with end points of the object, or the time between two events is determined by correlating them with an initial event in which the dial of a clock is at one position and a later event in which it is at a different position.[32] Our conception of physical space-time points is thought to arise from a correlation of points in two different (intuitive) sensory dimensions.

> If I look at my pencil from different sides, no one of the complexes of [subjective] elements that I experience is itself the pencil. The pencil is an object different from all these complexes: it is definitely "a thing in itself" in our sense. . . . [A]ll of these complexes . . . merely represent the object, that is, they are correlated with it. The details of their relation to it can be determined by physics and physiology only after the properties of the object are ascertained

[30] Ibid., 272.
[31] Schlick discusses the role of inference in moving from the identification of a point in the subjective order to a corresponding point in the objective order on pp. 274–75.
[32] Schlick discusses such measurements on p. 275.

> more closely, that is, only when we succeed . . . in designating it uniquely by means of general concepts.[33]

By "general concepts," Schlick means physically objective concepts, intersubjectively available to all because they abstract away from all private, phenomenal content. We are told, in effect, that we can't know anything about the relationship between the pencil and our private experiences that "represent it" until we have completed the construction and so can designate the pencil uniquely by physically objective contents alone.

He continues:

> If, while I am looking at the pencil, I touch its point with my finger, a singularity occurs simultaneously in my visual space and in my tactile space: a tactile sensation suddenly appears in my finger, and the visual perceptions of the finger and of the pencil suddenly have a spatial datum in common—the point of contact. These two experiences . . . are now correlated with the one and the same "point" of transcendent space, namely the point of contact of the two things "finger" and "pencil". The two experiences belong to different sensory domains and are in no way similar to one another. But what they do have in common is that they are singularities or discontinuities in what is otherwise a continuous field of perceptions surrounding them. It is through this feature that they are picked out from the field. This is how they can be related to one another and correlated with the same objective point in space.[34]

> The whole process of ordering things rests on effecting coincidences of this sort. Two objects are made to coincide . . . and this produces singularities inasmuch as the locations of these two otherwise separated elements are brought together. Thus, in the transcendent space-time schema, there is defined a system of distinct positions or discrete places that can be enlarged at will and extended in thought into a continuous manifold that permits the complete incorporation of all spatial objects.[35]

The terminus of Schlick's imagined "construction" is the physical world of objective space-time points that stand in various quantifiable relations to one another. This, for him, is the domain of objective knowledge.

Subjective awareness of our own sense experience is not part of this domain. Although conscious experiences are, of course, part of objective reality, we can have objective, communicable, interpersonal knowledge of them only after we have subsumed them under objective physical concepts by reducing them to brain processes.[36] What's more, the properties of which we take ourselves to be directly aware in sensation—colors, visual shapes, heat, cold, etc.—are private and incommunicable. These subjective

[33] Ibid., 272–73.
[34] Ibid., 273.
[35] Ibid., 274.
[36] Ibid., pp. 287–88.

properties are not real qualities of any element of objective reality.[37] They are, however, correlated with purely objective properties, which we discover scientifically and hence conceptualize.

But conceptualization is one thing, direct awareness is another.[38] In general, for Schlick, whatever we can be directly aware of cannot be objectively known, and whatever can be objectively known cannot be something of which we are directly aware.[39] Though the resulting system is ingenious, the strain in it is apparent. How can private, incommunicable, phenomenal content that cannot itself be objectively known be the epistemological foundation of a substantially realist conception of empirical science, and indeed, for everything that can be objectively known? This was a conundrum with which Schlick would struggle for more than a decade, finally allowing his earlier scientific realist side to be eclipsed by his growing attraction to verificationism.[40] It was also, as we shall see, a central problem in the *Aufbau* for the early Carnap, whose later embrace of verificationism, though different from Schlick's, reflected similar pressures.

5. THE KANTIAN LEGACY: CONTINUITY AND REACTION

According to Kant the truths of arithmetic are synthetic a priori because they are based on the pure a priori "intuition" of time, while the theorems of geometry are synthetic a priori because they are based on the pure a priori "intuition" of space. As explained in volume 1, Frege disagreed with the first of these claims. Using his new logic to reconceptualize the notion of analyticity and arguing that arithmetic is reducible to logic, he maintained that arithmetical truths are analytic.[41] Mach, Duhem, Schlick, and later Carnap, disagreed with Kant's second claim, holding that geometry is not an a priori science. Rather, they argued, its subject matter is neither "intuitively" given nor restricted to what we can visually imagine, but holistically determined by its place in our an overall physical theory.

Though the language of "construction"—as in "the construction of physical space"—continued to be used by philosophers like Mach, Schlick, and Carnap, the constructions in question were not Kantian shapings of the *reality of external appearance* by a *transcendental ego*. For one thing, the self was as much a construction for these philosophers as anything else; there was no unity of consciousness set off as the source of other "constructed" entities. For another thing, the "constructions" of the early

[37] Ibid., p. 279.
[38] Ibid., pp. 279–82.
[39] Ibid., pp. 285–86.
[40] This struggle is discussed in chapter 1 of Friedman (1999). See in particular pp. 41–42.
[41] His conception of analyticity is discussed in Soames (2014), pp. 41–42; his logicist reduction is discussed on pp. 45–59.

logical empiricists (here including the early Schlick and Carnap but excluding the phenomenalist Mach) did not accord "intuitions" (private sense experiences) robust epistemological priority. They were not entities our direct knowledge of which was the definitional base to which all other knowledge, e.g., of the physical world, was to be uniquely reduced. We have seen that for Schlick—who maintained that genuine knowledge of private experiences arises only after a reduction of the mental to the physical—this priority was supposed to be somehow reversed. Although it is not clear that Carnap went that far in the *Aufbau*, we will see that he too rejected the classical phenomenalist assignment of unique epistemological priority to claims about individual sense experiences.

Nevertheless, descendants of one well-known Kantian idea—sometimes expressed by the slogan "Concepts without percepts are blind; percepts without concepts are empty"—remained. For Schlick and Carnap objective knowledge of sense experiences—the contents of which are themselves private and incommunicable—requires abstraction to bring them under holistically interconnected concepts that are intersubjectively available to all. I have already mentioned the severe problem this idea posed for Schlick. In chapter 6, I will explain both the damage it did to Carnap and what might be salvaged from it. Before doing that, however, it is necessary to trace the powerful influence that Wittgenstein's tractarian philosophy of logic, language, and the modalities exerted on the rising leaders of logical empiricism.

6. THE IMPACT OF WITTGENSTEIN

After he published *The General Structure of Knowledge* in 1918, Schlick studied Russell's work in logic, which led him to abandon his earlier doctrine that all deductive reasoning is syllogistic in form. Shortly after he took up the Chair in Philosophy of the Inductive Sciences at the University of Vienna, he attended a seminar given by Hans Hahn in which he was introduced to the tractarian doctrine that logical truths are tautologies that "say nothing"—i.e., make no claim whatsoever—and hence constitute no threat to the idea that all knowledge is empirical. Taking this to be a major breakthrough, Schlick and his Vienna colleagues devoted two academic years to analyzing the *Tractatus*.[42] In December of 1924, he wrote to Wittgenstein expressing his own, and his colleagues', admiration of the work, his belief in the importance of spreading its doctrines, and his desire to meet the author.[43] Nevertheless, his first meeting with

[42] See p. xviii of Gordon Baker's preface to Wittgenstein and Waismann (2003).

[43] Schlick's letter is reproduced in part in Waismann (1979), p. 13. Weismann, who was closely allied with Schlick, is the member of the Circle who had the most contact with Wittgenstein, and who most thoroughly devoted himself to presenting and extending Wittgenstein's ideas. See, especially, Wittgenstein and Waismann (2003), including its historically informative preface.

Wittgenstein didn't occur until February of 1927, on one of the latter's visits to Vienna. After several private meetings with Wittgenstein, Schlick invited others, including Waismann, Carnap, and Feigl, to join them for conversations that occurred intermittently until Wittgenstein returned to Cambridge in January 1929. That move as well as Wittgenstein's chilly reaction to the positivists' 1929 manifesto "The Scientific Conception of the World," and various practical difficulties, limited Wittgenstein's contact with Schlick and with other members of his circle through 1936, when Schlick died.[44]

The impact of the *Tractatus* on the thinking of these early members of the Vienna Circle was profound and, from the distance of nearly a century later, somewhat surprising. It is not surprising that the anti-metaphysical *Tractatus* reinforced the already strong anti-metaphysical tendencies of Schlick and Carnap. It is more surprising to learn that it helped move them away from scientific realism and toward both phenomenalism and verificationism. There are, to be sure, notable verificationist themes in the *Tractatus*—including the non-cognitive treatments of value and the meaning of life, the denial that there are genuine but unsolvable problems, or meaningful but unanswerable questions, the dismissal of philosophical theses as violations of the tractarian criterion of intelligibility, the renunciation of truth as a philosophical goal and its replacement by the goal of the dissolution of linguistic confusion. However, neither phenomenalism nor the repudiation of scientific realism leap from the pages of the *Tractatus*. Although a few passages deal with the interpretation of scientific matters, they don't seem to add up to an explicit endorsement of anti-realism. Nor do Wittgenstein's metaphysical simples, which are, in their way, the subject matter of all meaningful tractarian propositions, plausible candidates for phenomenal sense data or sense experiences.

Nevertheless, four tractarian doctrines did conspire to help push Schlick and others toward the combination of phenomenalistic verificationism with scientific anti-realism that was to become closely associated with logical empiricism:

(i) All epistemic and metaphysical modalities are linguistic, and ultimately logical, modalities.
(ii) Since all meaningful sentences are truth functions of atomic sentences, the truth values of all meaningful sentences are settled by the truth values of atomic sentences.
(iii) An atomic sentence S is true (false) iff the objects $o_1 \ldots o_n$ designated by its names stand (don't stand) in the relation R in which they are represented as standing by the linguistic relation in which the names in S stand to one another. This will be so iff there is (isn't) an atomic fact consisting

[44] Waisman (1979), pp. 18–27.

of $o_1 \ldots o_n$ standing in R. Hence to know that S is true (false) is to know that $o_1 \ldots o_n$ stand (don't stand) in R.

(iv) Reality is the totality of atomic facts.

Imagine yourself in the shoes of the author of *The General Theory of Knowledge,* confronted with these tractarian doctrines. For you, physical space-time points plus objects occupying them, and the events occurring there, are not primitive tractarian metaphysical simples but "constructions." These constructions are the entities designated, described, or quantified over by physical theory. When generality is treated truth-functionally, as it is in the *Tractatus,* you (standing in Schlick's shoes) are willing to take all statements of physical theory to be truth functions of what seem to be atomic statements about physical objects, events, and space-time points. But the *Tractatus* has reinforced your conception that the process of analysis does not stop there. The properties and relations predicated of physical objects, events, and space-time points by the pseudo-atomic statements of physical theory are, as you have emphasized, conceptually interdependent and holistically understood. Because these statements bear conceptual relations to one another, they are *not* independent in the sense that atomic statements are required (by the *Tractatus)* to be. Genuine atomic statements, which are, of course, *logically independent* of one another, have to be epistemically and metaphysically independent if, according to the fundamental atomist doctrine (i), relations of logical dependency are to replace all conceptual relations of epistemic or metaphysical dependency.

For this replacement to occur, all pseudo-atomic statements of physical theory must be understood to be truth functions (in the *tractarian* sense) of genuine atomic statements, the truth or falsity of which are independent of each other. Once this level is reached, one can determine the truth of each atomic statement *independently* of assumptions about any other statements. What might the subject matter of such statements be? When atomic statements are conceived to be radically independent in this way, it is quite natural to think of their subject matter as nothing more than the momentary sense impressions of an agent whose apprehension of the sense data named by the constituents of an atomic statement is simultaneously the verification of that statement and the agent's understanding of it.

Think of the atomic statement along the lines of a use of *This is P* where 'this' designates a momentary sense datum d and 'P' is replaced by a predicate expressing a phenomenal property about which one cannot be mistaken. One can't apprehend the statement until d is perceived, at which point one will immediately know whether it is true or false—*without having to rely on any assumptions about other atomic statements.* With this the journey from *The General Theory of Knowledge* to phenomenalistic anti-scientific realism, by way of the *Tractatus,* is complete.[45]

[45] This sense of independence, capturable by Schlick, approximates but does not quite reach genuine *tractarian independence* of atomic propositions— if, as is natural to suppose,

Although Schlick did make this journey, he seems to have done so with some wavering back and forth. By the early 1930s, however, he had reached the end of the journey. In 1934, writing in "On the Foundation of Knowledge," he declares that philosophy seeks "an unshakable, indubitable, foundation a firm basis on which the uncertain structure of our knowledge could rest."[46] The foundation is described as "the natural bedrock which exists before all building and does not itself totter."[47] The bedrock consists of true sense-data statements that are knowable merely by understanding them. Calling them "confirmations," he characterizes them as the only *synthetic* statements in which understanding and knowledge of truth coincide. He says,

> I can understand the sense of a "confirmation" only by, and when, comparing it with the facts, thus carrying out that process that is necessary for the verification of all synthetic statements. While in the case of all other synthetic statements determining the meaning is separate from, distinguishable from, determining the truth, in the case of observation statements the two coincide—just as in the case of analytic statements. However different therefore "confirmations" are from analytic statements, they have in common that the occasion of understanding them is at the same time that of verifying them: I grasp their meaning at the same time as I grasp their truth. In the case of a confirmation it makes as little sense to ask whether I might be deceived regarding its truth as in the case of a tautology. [As if we couldn't be mistaken about a tautology being true while understanding it.] Both are absolutely valid. However, while the analytic, tautological, statement is empty of content, the observation statement supplies us the satisfaction of genuine knowledge of reality."[48]

Since error is impossible here, and all empirical knowledge is justified by our certain knowledge of the sensory given, we have epistemic foundationalism. Schlick had moved from a form of epistemic holism in 1918 to classic epistemic foundationalism in 1934.

There is no doubt that his reading of the *Tractatus* played a central role in the transformation. Speaking of the Vienna Circle in 1926, when Schlick brought Carnap to Vienna, Michael Friedman says:

> [T]he Circle understood the *Tractatus* as articulating a foundationalist-empiricist conception of *meaning*. Definitions explain the meanings of words in terms of other words, but this procedure cannot go on to infinity, or else no word ultimately has meaning at all. Therefore, all meaning must finally

---

predicates like 'is red' and 'is green' are logically simple constituents of atomic sentences. Since these are obviously and transparently mutually exclusive, while bearing no *logical* relation to one another, they violate the letter, but perhaps not the spirit, of the tractarian thesis (i) above.

[46] Schlick (1934 [1959]), p. 209.

[47] Ibid., p. 370.

[48] Ibid., p. 225.

> rest on primitive acts of ostension, and what is ostended must be immediately given.[49]

Friedman cites Viktor Kraft—a member of the Vienna Circle listed in its 1929 manifesto—as tying this conception of meaning to phenomenalist epistemology.

> Definitions are ultimately reducible to ostension of what is designated. One can point only at something which is immediately given, and thus only at what is perceivable. In this way, what assertions can possibly mean is tied to experience. No meaning can be given to that which is not reducible to experience; and this is a consequence of fundamental importance.[50]

Friedman concludes that "there is no doubt that this conception of meaning—and this understanding of the *Tractatus*—was adopted especially by Waismann and Schlick."[51] In this, he agrees with Kraft himself, who took this view as coming from the *Tractatus*.

> Wittgenstein identified [atomic propositions] with the propositions he called "elementary propositions." They are propositions which can be immediately compared with reality, i.e. with the data of experience. Such propositions must exist, for otherwise language would be unrelated to reality. All propositions which are not themselves elementary propositions are necessarily truth functions of elementary propositions. Hence all empirical propositions must be reducible to propositions about the given.[52]

Although the other members of the Vienna Circle were certainly aware of this reading of the *Tractatus* and of the path that led from it to phenomenalistic verificationism, they didn't all follow Schlick down it. Otto Neurath became the most notable dissenter. However, most members of the Circle, including Carnap, were at least influenced by the position, even if they didn't fully or consistently endorse it. In the next chapter, I will discuss Carnap's most important early work, *Der logische Aufbau der Welt—The Logical Structure of the World*.[53] Although published in 1928, two years after he arrived in Vienna, the initial manuscript was completed in 1925, a year before he arrived. In it we find early versions of major themes that were to occupy him for decades.

[49] Page 148 of the reprinting of Friedman (1992) in Friedman (1999).
[50] Kraft (1950), pp. 32–33.
[51] Friedman (1999), p. 148. As Friedman notes, confirmation of the point about Waismann is found in section 7 of Waismann (1979). As Friedman also notes, Schlick adopted this position consistently starting in the early 1930s.
[52] Kraft (1950), p. 117.
[53] Carnap (1928 [1967]).

CHAPTER 6

Carnap's *Aufbau*

1. THE STRUCTURE, GOALS, AND VARIETY OF CARNAPIAN REDUCTIONS

The goal of the *Aufbau* is to establish the possibility of constructing a system that brings together all scientific knowledge in a single reductive conceptual framework in which concepts sufficient for all of science are defined from a small base of primitive concepts and all claims expressing genuine scientific knowledge are translated into claims involving only logical concepts plus (perhaps) the primitives. That Carnap didn't attempt to articulate more than a tiny fraction of any such reduction isn't really a shortcoming; it is unlikely that anyone will ever do much more. His aim was the more modest one of establishing that such a system is *possible*. Doing so would, he believed, demonstrate *the unity of science* by showing that all scientific knowledge can be conceptualized as knowledge of a single domain of objects bearing the primitive properties and relations of a reductive constructional system.

In one respect the constraints on reduction were strict. Carnapian reductions required *definitions* of expressions to be eliminated that correlate each formula f containing such an expression with an extensionally equivalent formula g in which the expression does not occur.[1] In another respect the constraints were quite loose. Because only *extensional equivalence* was required, no modal, epistemic, or explanatory conditions were imposed. This allows for significant conceptual *revision,* because Carnapian *definitions* were taken to be sufficient to eliminate the defined expressions from theorems of the theory, and hence from the explication of the scientific knowledge provided by the theory.[2]

This feature of Carnap's project gives rise to two general questions. (i) Is there any reason to suppose in advance that our scientific knowledge *can be* revised, explicated, and unified in a reductive conceptual system of this sort? Does the mere fact that we now possess (some) scientific knowledge of various domains *guarantee* that the theories expressing that knowledge *must be* reducible to a theory of a single domain, knowledge of which explicates the scientific knowledge we now have? (ii) Is there any reason to suppose that *if* our theory of one scientific domain *is*, in principle, reducible *in Carnap's formal sense* to a theory of the primitive properties and relations born by the elements of an underlying domain, then a successful reduction *will show* that we are capable of knowing the former theory by knowing the latter? Carnap seems to have taken it for granted that the answers to these questions are in the affirmative. This, I will argue, is questionable.

In the *Aufbau,* Carnap insists that several reductions are theoretically possible. Three reductions of all (possible) scientific knowledge to knowledge of physical facts are mentioned in the text. A fourth possible physical reduction is added in the preface (written in 1961) to the second edition published in 1967. The three possible but sketched physical reductions mentioned in the text are:

(i) A physicalistic reduction that takes electrons standing in certain primitive spatiotemporal relations to be the fundamental objects. Properties of electromagnetic fields are said to be definable in terms of acceleration of electrons, atoms are defined, and gravity is said to be definable in terms of acceleration of atoms. All other physical things are ultimately to be reduced to magnetic fields, electrons, and gravitation. Since all things are, at bottom, physical, all psychological knowledge and even all cultural knowledge is said to be reducible, in principle, to knowledge of the

[1] See sections 48 and 49 of the *Aufbau* for the need for exceptionless universal generalizations as definitions.

[2] In 1961, thirty-three years after the publication (in German) of the *Aufbau*, Carnap recants both the insistence on definitions and the failure to impose intensional constraints on reductions. See Carnap (1928 [1967]), pp. viii–x of the 1961 preface to the second edition.

physical, and ultimately to the spatiotemporal relations in which electrons stand to one another.[3]

(ii) A physicalistic reduction to points of four-dimensional space-time standing in relative location relations plus relations between these points and real numbers representing "potentials."[4]

(iii) A reduction to Minkowski's "world-lines."[5]

While each of these imagined reductions requires one to "construct" everyday physical objects, human bodies, brains, and neurological events, the reduction imagined in 1961 envisions reducing all scientific knowledge to knowledge of everyday physical objects bearing observable properties and standing in observable relations to one another.[6]

To deal with the relationship between the psychological and the physical, it is necessary in each of these systems to establish correlations between neural events and (reported) thoughts, feelings, sensations, and the like, with the goal of correlating every type of psychological event or state with a corresponding type of neurological event—so that each instance of the neurological type is correlated with an instance of the corresponding psychological type. This is supposed to it make it possible to formulate a true universally quantified biconditional that "defines" each psychological type in terms of a neurological type, which, in turn allows one to *replace* all psychological language with physical language, thereby completing the reduction of the psychological to the physical.[7] A further reduction of the cultural to the psychological is envisioned.

Although Carnap asserts the possibility of the physicalistic reductions in the *Aufbau,* they don't play a large role in the work. They are mentioned in passing in order to shed light on the reduction he is most concerned with, which is phenomenalistic, or psychological. Two types of psychological reduction are said to be theoretically possible. One starts from an *autopsychological base*, the elements of which are undifferentiated experiences of a single subject. These are short, temporally extended cross sections of experience that may involve any of the individual modes of sense perception—vision, touch, hearing, etc.—or any simultaneous combination of them. The only primitive concept applying to these basic elements appealed to in the reduction is the relation *recollected similarity.*[8] Carnap's

[3] Ibid., p. 99. (i)–(iii) are paraphrases, not quotes.
[4] Ibid., p. 99.
[5] Ibid., p. 100.
[6] Ibid., pp. vii–viii of the introduction to the second edition.
[7] Ibid., p. 92.
[8] On page vii of his 1961 preface to the second edition, Carnap says, "I should now prefer to use a larger number of basic concepts, especially since this would avoid some drawbacks which appear in the construction of the sense qualities . . . I should now consider for use as basic elements, not elementary experiences . . . but something similar to Mach's elements, e.g., concrete sense data, as, for example, 'a red of a certain type at a certain visual field place

methodology is to use this relation to extract phenomenal concepts the extensions of which are classes of basic undifferentiated experiences known as "the given." These phenomenal concepts are imagined as providing the basis for constructing a series of increasingly sophisticated definitions resulting, as incredible as it may sound, in definitions of all objects of our knowledge.

The other envisioned phenomenalistic reduction is called "the general psychological reduction." It too starts from a domain of undifferentiated experiences as elements, only this time the base includes experiences of all subjects. In both reductions the physical is supposed to be reduced to the psychological, although in the autopsychological reduction human brains and bodies other than one's own are first "defined" in terms of the experiences of what will turn out to be the single subject that one is. After that, experiences of other subjects, and then those subjects themselves, will be defined in terms of the brains and bodies just defined. The remainder of the physical is then supposed to be reduced to the psychological. No matter which form of psychological reduction is chosen, Carnap took it to be possible to translate statements about physical objects into statements about psychological objects, and ultimately into statements about undifferentiated experiences standing in relation to one another.

2. ARE THE REDUCTIONS POSSIBLE?

Since neither Carnap nor anyone else dreamed of actually completing any reductions mentioned in the *Aufbau,* it is important to ask whether such reductions really are possible, and why it is supposed to matter whether they are. Carnap doesn't say a great deal about this, but he does hint at his reason for believing the autopsychological reduction to be possible.

2.1. Knowledge and Epistemic Primacy

Consider the following remarkable passage from section 57.

> Statements about physical objects can be transformed into statements about perceptions (i.e., about psychological objects). For example, the statement that a certain body is red is transformed into a very complicated statement which says roughly that, under certain circumstances, a certain sensation of the visual sense ("red") occurs. Statements about physical objects which are not immediately about sensory qualities can be reduced to statements that are.

---

at a given time.' I would then choose as basic concepts some of the relations between such elements, for example 'x is earlier than y', the relation of spatial proximity in the visual field and in other sensory fields, and the relation of qualitative similarity, e.g., color similarity."

> *If a physical object were irreducible to sensory qualities and thus to psychological objects, this would mean that there are no perceptible indicators for it. Statements about it would be suspended in the void; in science at least, there would be no room for it.* Thus all physical objects are reducible to psychological ones.[9]

How should we understand this? Presumably the conclusion should be qualified. *All statements about physical objects objectively known at a given time to be true can be replaced, without change of truth value, by translations that speak only of psychological objects.* Remember, Carnap thought of translation as proceeding by steps each of which involves using "definitions" to replace each physical-object formula PHYO with a formula PSYCO in psychological language, where the universal closure of the biconditional connecting the two formulas is a true sentence. In order to *know* that any given reduction is successful, one must, of course, *know* that every such "definition" used in the reduction is true. But this is not required for the physical to be *reducible* to the psychological. In order for that to be so, it is sufficient that the required definitions be *true*. Carnap didn't claim to know of any purported reduction that it is successful, but only that *there must be* a successful reduction of the physical to the psychological.

Why? He thought the physical must be reducible to the psychological because if it weren't, we wouldn't have the knowledge of the physical that we in fact have. He thought that we *recognize* and *come to know of* physical things by *recognizing* and *coming know about* our sense experience. This is clear from the very next section of the *Aufbau* after the passage just cited.

> We now have to decide whether our system form requires a construction of the psychological objects from the physical objects or vice versa. Because of their mutual reducibility, it is logically possible to do either. Hence, we have to investigate the epistemic relation between these two object types. It turns out that psychological processes of other subjects can be recognized only through the mediation of physical objects. . . . On the other hand, the recognition of our own psychological processes does not need to be mediated through the recognition of physical objects, but takes place directly. Thus, in order to arrange psychological and physical objects in the constructional system according to their epistemic relation, we have to split the domain of psychological objects into two parts: we separate the *heteropsychological* objects from the *autopsychological* objects. The autopsychological objects are epistemically primary to the physical objects [i.e., the latter are recognized and known by recognizing and knowing the former], while the heteropsychological objects are secondary. . . . Thus the sequence with respect to epistemic primacy of the four most important object domains is: the autopsychological, the physical, the heteropsychological, and the cultural.[10]

[9] Carnap (1928 [1967]), p. 92, my emphasis.
[10] Ibid., pp. 93–94.

Carnap appears to believe that our *evidence* for claims about physical objects is, or results from, our knowledge of our own mental states, while our evidence for claims about the psychological states of others is, or results from, our knowledge of certain physical things. *So, he thinks, knowledge of our own mental states provides all our evidence for any knowledge we have of propositions about the world.* Suppose, for the sake of argument, this is right. Given this, we next consider the possibility that there are no true, universally generalized biconditionals connecting formulas about our sensory experiences with various physical-object formulas we ordinarily take ourselves to know on the basis of those experiences. What, if anything, might we then conclude? Without such universal generalizations, Carnap would, I suspect, conclude that physical-object statements previously thought to be known would, in fact, *not* be known—either because they would be false (even if the statements expressing our sensory evidence for them were true) or they would be true but insufficiently supported by our evidence. So, he would argue, without exceptionless correlations between the psychological and the physical, we wouldn't know statements we in fact do know. They would, as he vividly puts it, "be suspended in a void." Since we *do* know the relevant physical-object statements, reducibility *must* be possible. That, I believe, was the source of his confidence in the reducibility of the physical to the autopsychological.

2.2. The False Guarantee of Reducibility

This justification of Carnapian confidence is unconvincing. Think of the vast range of potential knowledge to be covered by any proposed "reduction" of the physical to the psychological. If the aim is to "unify science," then the statements to be "reduced" to extensional statements about one's own sense experiences must include those of theoretical physics, including those reporting the behavior of what we take to be the most fundamental physical objects—subatomic particles, say—throughout the universe. Surely it is impossible to reduce all these statements to statements about one's own sense experiences; the reductive base of sense experience is too meager. The point would hold even if the base were expanded to include the actual sense experiences of every human agent, or even all observational statements about the everyday physical objects any human agents have ever or will ever perceive. When the domain to be reduced is so much richer than the domain to which it is to be reduced, no significant reduction is possible, unless either (i) we eliminate from the domain to be reduced all statements not definitely known to be true, or (ii) we take a nonrealistic view of the statements to be reduced, assuming in advance that when two such statements can't be distinguished in terms of the reductive base, they must either be excluded from the reduction or identified as two formulations of some other reducible statement.

Neither (i) nor (ii) fits Carnap's project very well. As he would surely agree, scientific inquiry is fluid, consisting at any one time of a limited

amount of what we know plus much more in which we have shifting degrees of confidence and varying levels of justification. If *science* is to be unified by a Carnapian constructional system, none of the statements that are not yet definitely known but are, nevertheless, scientifically in play at a given time can be excluded. All must be represented in the system, thereby ruling out strategy (i) for securing reducibility. As for strategy (ii), let S1 and S2 be inconsistent statements about subatomic particles spatiotemporally removed from us, about which we have no evidence that allows us to confidently decide between them. This alone is no basis for taking them to have the same truth value (S2 might be the negation of S1). Nor is it enough to be indifferent about adding S1 to our currently accepted body of scientific statements versus adding S2. Thus, we have a problem. Although Carnapian constructional systems that attempt to reduce the physical to the autopsychological, to the general psychological, or even to the ordinarily physically observable, may need to discriminate S1 from S2, we have been given no explanation of how this should be done. One might consider pursuing parallel reductions in some cases, but that practice couldn't be followed very long without generating far too many options.

2.3. Phenomenalist Temptation vs. Metaphysical Neutrality of Carnap's Reductions

What, then, explains Carnap's seemingly unquestioning confidence in reducibility? It must, I think, be an implicit way of assigning objective empirical content to theories. Classical phenomenalism, exemplified by Russell (1914b) and (1918–19), illustrates what is needed. The classical phenomenalist starts from reports of sense experiences, which are taken to be unproblematically meaningful and capable of being known to be true (or false). Ordinary physical-object statements are taken to be definable from these, while the theoretical statements of physics are definable in similar but more complicated ways. When the classical phenomenalist says that everything is so *definable*, he is identifying the (knowable) content of physical-object statements—and hence what one knows when knows them to be true—with the contents of statements explicitly about sense experiences. On this view, *all knowledge is knowledge of sense experiences and nothing else*. When theoretical reduction is conceived in this way, pretheoretic claims that aren't experientially definable are *dismissed*, not simply as being currently unknown, but as being either *unknowable* in principle (Russell 1914b, 1918–19) or, in coming formulations of logical empiricism, as failing to have any empirical meaning at all. On this latter, more ambitious view, the very contents (meanings) of individual physical-object statements are identified with the contents of extremely complex but definitionally equivalent statements that speak only of sense experiences. Consequently, an inventory of the world that mentioned each agent and each experience but nothing else would leave nothing out.

On either the more modest Russellian form of phenomenalism or the more ambitious logical empiricist form, pretheoretic statements that resist reduction are simply excluded as not capable of contributing to human knowledge. On the surface, it might seem that Carnap's autopsychological reduction could, in principle, be understood in either of these two ways. According to one, it is decided in advance that the domain of the knowable is one's own sense experience. According to the other, the range of meanings of one's statements is confined to claims about one's own sense experience. Either way, there can be no worries about the reducibility of any genuinely meaningful claims that can be known. They must be reducible, because one has decreed them to be so at the outset.

Although this interpretation of the *Aufbau* allows one to explain Carnap's bafflingly breezy confidence that the physical must be reducible to the psychological, it can't be correct. The interpretation doesn't explain either his equally breezy confidence in the reducibility of the psychological to the physically fundamental, or the relationship he took to hold between the autopsychological and the general psychological reductions. *In a classically phenomenalist reduction, the base to which all the other levels are reduced provides all the knowable, or all the meaningful, content of every statement*. This can't have been Carnap's conception of reduction in the *Aufbau*. If it had been, then accepting the autopsychological reduction would have involved either believing that other people *might* exist, but one could never know whether they do (because the only contents one can know concern solely one's own sense experiences), or believing that to say or think that other people exist is simply to say or think that one has certain sense experiences oneself. Carnap didn't subscribe to these absurdities. Nor can the envisioned conception of Carnapian reduction—as replacing realist readings of non-base statements with readings in which their contents are given by complex base statements—explain how he could regard all his envisioned reductions as equally correct and noncompeting.

Thus, we still don't have satisfactory answers to the two most significant interpretive questions about the *Aufbau*.

Q1. Why was Carnap so confident that different constructional systems "reducing" the knowable world to very different conceptual bases *must* be possible?

Q2. Why was he confident that these different "reductions" are equally correct and noncompeting?

Here is a sample of relevant passages.

> We now have to decide whether our system form requires a construction of the psychological objects from the physical objects or vice versa. Because of their mutual reducibility, it is logically possible to do either.[11]

[11] Ibid., p. 93.

> If it is not required that the order of construction reflect the epistemic order of objects, other systems are also possible. . . . Since all cultural objects are reducible to psychological, and all psychological to physical objects, the basis of the system can be placed within the domain of physical objects. Such a system form could be called *materialistic*. . . . However, it is important to separate clearly the logico-constructional aspect of the theory from its metaphysical aspect. From the logical viewpoint of construction theory, no objection can be made against scientific materialism. Its claim, namely, that all psychological (and other) objects are reducible to physical objects is justified. Construction theory and, more generally, (rational) science neither maintain nor deny the additional claim of metaphysical materialism that all psychological processes are essentially physical, and that nothing but the physical exists. The expressions "essence" and "exists" (as they are used here) have no place in the constructional system, and this alone shows them to be metaphysical.[12]

> The [pretheoretic] realistic language, which the empirical sciences generally use, and the constructional language have actually the same meaning: they are both neutral as far as the decision of the metaphysical problem of reality between realism and idealism is concerned. . . . *Let us emphasize again the neutrality especially of the constructional language*. This language is not intended to express any of the so-called epistemological, but in reality metaphysical, doctrines (for example, realism, idealism, solipsism), but only epistemic-logical relations. In the same sense, the expression "quasi object" [Carnap's term for types of objects defined in constructional systems] designates only a certain logical relationship and is not meant as the denial of a metaphysical reality. It must be noted that all real objects (and constructional theory considers them as real to the same degree as do the empirical sciences) are quasi objects. Once it is acknowledged that the realistic and the constructional languages have the same meaning, it follows that constructional definitions and the statements of the constructional system can be formed by translating . . . statements which are found in the realistic language of the empirical sciences. *Once realistic and constructional languages are recognized as nothing but two different languages which express the same state of affairs, several, perhaps even most, epistemological disputes become pointless.*[13]

The main points expressed here are (i) that various ways of unifying science by reducing all objectively knowable statements to markedly different conceptual bases are possible, (ii) that scientific theories expressed in terms of these unifications are equally correct because they *stand for the same states of affairs* and *have the same empirical*—i.e., non-metaphysical—*meaning*, (iii) that the choice of a particular constructional system for unifying science involves no metaphysical commitments involving such

[12] Ibid., pp. 94–95.
[13] Ibid., pp. 86–87, my emphasis.

doctrines as realism versus idealism, and (iv) that such traditional metaphysical disputes are pointless, and indeed may well be empirically meaningless. The crucial issue needed to answer Q2 is raised in (ii). In order to compare different constructional systems for unifying science, one needs an external benchmark against which each can be tested. To understand what it means to say that different unifications *stand for the same state of affairs* or *have the same empirical meaning,* we must understand what this benchmark amounts to. Because Carnap is nearly silent about this point, it is up to us to fill in the needed content in a way that can be made consistent with the totality of his remarks.

3. CAN THE *AUFBAU* BE MADE COHERENT?

One way to make the *Aufbau* coherent is to take the evidential base for objective empirical knowledge to consist of all possible sense experiences of human subjects. This evidential base is the class of potential observational data against which theories are to be tested. Making this decision requires using a notion of *possible experience* that goes beyond evidence or experience that can't *logically* be ruled out, and also beyond experience that can't be ruled out by *a priori* reasoning alone. The possible sense experiences required for this conception of observational data are *not* those described by any *logically consistent*, i.e., noncontradictory, sets of sentences about our experience; nor are they those described by any collection of propositions about our experience not knowable *a priori* to be false. What is needed are experiences human subjects are *capable* of having, perhaps those that are, as some today might say, *metaphysically* possible for us to have. This is not a notion Carnap officially recognized, but it is one he needed.

Next, we identify the meaning, or knowable empirical content, of a unification of science expressed by a constructional system with the class of possible sense experiences of any and all agents with which it is compatible. On this interpretation, the *Aufbau* implicitly endorses a phenomenalistic version of holistic verificationism. According to this view, it is scientific systems as wholes that have empirical meaning or content. Consequently, two systems with different primitive bases employing their own "definitions" of Carnapian "quasi-objects" at various theoretical levels have the same content, and so express the same potential human knowledge, if and only if they fit the same possible sensory experience. In calling the objects posited by a theory "quasi objects," Carnap signals that reductions to different primitive bases generated by theory-internal definitions do *not* result in different ontologies—e.g., materialism versus idealism. To think otherwise is to misunderstand the relationship between the theory and the reality it describes. Non-observational statements of a theory do *not* directly stand for any elements of reality; they merely contribute to the empirical content of the theory as a whole, which is the totality of its predictions about possible

sense experience. Although Carnap doesn't explicitly acknowledge this way of looking at things, it provides him with what he needs.

Moreover, it's not the only way of doing so. A different version of holistic verificationism is possible in which the meaning or empirical content of a particular unification of science is given by the *intersubjectively observable events* predicted by the unified constructional system as a whole. What Carnap required to secure the metaphysical neutrality of his different imagined constructions was a common denominator involving observational predictions needed to assess them. Although he did, when writing the *Aufbau,* think of perception and observation phenomenalistically, he didn't have to. Any notion of observation, and hence empirical content, would do, provided that it could be utilized no matter which reductive base—autopsychological, heteropsychological, or physicalistic—was chosen. In principle, either the possible sensory experiences of arbitrary human agents or the physical events observable by possible human beings could play this role.

Next we consider Carnapian definitions, which, he thought, were required to connect non-observational claims with observational claims. The Carnap of the *Aufbau* seemed to think of theories along the lines of a certain restricted version of the hypothetical-deductive model. On this conception, theoretical statements not containing observational vocabulary, sometimes together with observational statements, make observational predictions by *logically entailing* further observational statements. If these further statements are true, the theory is partially confirmed; if they are false it is disconfirmed. When one thinks of the relationship between theory and evidence this way, in terms of *logical* consequence, *definitions* of the non-observational vocabulary in terms of the observational vocabulary—thought of as *conventions* that don't themselves have to be empirically verified—may seem to be mandatory, if the theory is to make any predictions (and hence have any empirical content) at all. Since Carnap had no doubt that science does make many testable predictions, he had no doubt when he wrote the *Aufbau* that definitions of the sort he took to be required *must* be possible.

In later years he came to realize that there is no need for the connection between theoretical hypotheses and observational predictions to be so tightly constrained. Although the non-observational parts of a theory must be connected with the observational parts, the connection need not be made by *definitions.* For the theory to logically entail observational consequences it is sufficient that it contain universally quantified conditionals (rather than biconditionals) the antecedents of which contain theoretical vocabulary and the consequents of which contain observational vocabulary. Not having the epistemic status of definitions that *replace* one set of concepts with another, these bridge principles are just more theory—auxiliary hypotheses needed to endow the more abstract parts of the theory with empirical content.

This is what Carnap was talking about when he said the following in his 1961 preface to the second edition.

> One of the most important changes [from the position taken in 1928] is the realization that the reduction of higher order concepts to lower level ones cannot always take the form of explicit definitions; generally more liberal forms of concept introduction must be used. . . . The positivist thesis of reducibility of thing concepts to autopsychological concepts remains valid, but the assertion that the former can be defined in terms of the latter must now be given up and hence also the assertion that all statements about things can be translated into statements about sense data. Analogous considerations hold for the physicalist thesis of reducibility of scientific concepts to thing concepts and the reducibility of heteropsychological concepts to thing concepts. . . . [In 1956] I considered a method which was already used in science . . . namely the introduction of "theoretical concepts" through theoretical postulates and correspondence rules. . . . The correspondence rules connect the theoretical terms with observational terms. Thus the theoretical terms are interpreted, but this interpretation is always incomplete. Herein lies the essential difference between theoretical terms and explicitly defined terms. The concepts of theoretical physics and other advanced sciences are best envisioned in this way. At present I am inclined to think that the same holds true of all concepts referring to heteropsychological objects whether they occur in scientific psychology or in daily life.[14]

Finally, we need to understand the significance Carnap attached to the autopsychological reduction. First, he took it to explain how each individual's knowledge, not only of theoretically foundational physical objects, but also of non-fundamental physical objects, other persons, and their sense experience, is grounded in the individual's own sense experience. To say that it is so grounded is *not* to say that the *content* of the autopsychological construction of science is restricted to the individual's own sense experience. It had better not be. As with all constructions, *the content* of the unified autopsychological system of science is the set of observable predictions it makes—either about the possible sense experience of human agents or about intersubjectively observable physical events. Crucially, however, Carnap thought that the extent to which any individual agent does *know* this content is the extent to which *the agent's own sense experience* justifies believing those observational truths.

Second, the autopsychological reduction is seen by Carnap as providing a way of abstracting general content—graspable by any agent—from the private, idiosyncratic, sensory content of an individual agent. It is this abstracted content that is needed when characterizing the contents of all Carnapian reductions either in terms of possible sensory experience or

[14] Ibid., pp. viii–ix.

in terms of intersubjectively observable events. As we will see in section 3.5, Carnap thought that objective knowledge shared by different agents cannot include the phenomenal contents of any particular sense experiences. His strategy was to eliminate reference to any such particularized contents by identifying the place particular types of sense experience occupy in the sensory systems common to human beings—visual, auditory, tactile, etc.

For example, when I have a phenomenally red sense datum, I have a visual experience that stands in various abstract relationships to other visual experiences of mine, and to my experiences arising from other sense modalities as well. Call a visual experience that stands in these relationships to my other experiences one of my R-experiences. Recognizing the impossibility of comparing my phenomenally red sense datum with anyone else's sense datum, Carnap plausibly maintained that there is no such thing as objective—i.e., sharable—knowledge of phenomenal content. But he did seem to think that different agents could have R-experiences. It was sensory experience in this sense—with specific phenomenal contents abstracted away—that he took to be capable of being intersubjectively known, and thus to provide the ultimate contents of all human knowledge. This abstraction is one of his chief concerns in setting out the framework for the autopsychological reduction.

4. SHARED WORRIES FOR ALL REDUCTIONS

Having attempted to make Carnap's conception of multiple noncompeting unifications of science coherent, we need to address remaining problems shared by all his attempts at unification. One problem for the constructional systems envisioned in the *Aufbau* was the conception of reduction by definition, which requires the truth of universally quantified biconditionals. I have already explained why Carnap assumed that the ability of one's evidence to underwrite one's theoretical knowledge requires exceptionless correlations between theoretical and observational vocabulary. Thinking that verification of theoretical claims requires assessing the observational claims they *logically entail,* and taking this entailment to require one's theoretical vocabulary to be *definable* in terms of one's observational vocabulary, the Carnap of the *Aufbau* had, in effect, an implicit "transcendental argument for definitions." *Since without definitions theories we know to be testable wouldn't be testable, there must be definitions.* Of course, this argument was misguided. Although one needs principles connecting the non-observational to the observational, the principles don't have to be definitions, a priori truths, or even universally quantified biconditionals. As already noted, they can be just more theory. So conceived, there may be no way of verifying (or falsifying) them independently of verifying (or falsifying) other parts of the total theory. But, as we now realize, this doesn't

distinguish them from many other statements. The Carnap of the *Aufbau* didn't realize this, hence his emphasis on definitions.

Given this, one should not be surprised that his reliance on definitions should cause problems. Think again about the imagined reduction of the physical to the autopsychological. Couldn't one's perception of red things be generally reliable, and so lead to knowledge, even if exceptions sometimes occurred—in which what looks red isn't, or what is red isn't seen as red? Surely it could, even if one restricts oneself to knowledge of the color of things one is looking at in good conditions. But that is only the beginning. We also know the color of many things we aren't currently looking at, as well as the color of some things no one has ever seen, but would be perceived as having a specific color *if one were ever to look at them*. There is no reason to assume that true, universally generalized biconditionals are required in all such cases.

The problem is exacerbated by Carnap's uncritical attitude toward the concept of *knowledge* when writing the *Aufbau*. Although his goal was to provide a unified conception of the scientifically knowable, the book contains no sustained examination of what knowledge is. One can get a sense of what he missed by considering a conception of knowledge that *wouldn't* vindicate his presupposition that knowledge-guaranteeing definitions must be possible. For this purpose we may accept a Williamsonian analysis of knowledge as *safety* plus a dubious Carnapian premise about the evidential role of knowledge of one's own sensory experiences.[15]

Let p be a true physical-object statement I believe on the basis of certain sense experience. Three troubling possibilities present themselves. (i) Perhaps there is a true, exceptionless universal generalization UG of the sort Carnap imagined that connects statements about my sense experience with physical-object statements like p. But UG may be a mere accidental generalization that doesn't support counterfactuals. If so, then although p may be true and believed by me, it might also be true that I *could* rather easily have been in my present state of accepting p, even though p was false. If so, I wouldn't know p, even though the physical-object statements I take myself to know are reducible-in-the-Carnapian-way to statements about my sense experiences. Thus, even a successful Carnapian "reduction" wouldn't explicate my knowledge. (ii) Perhaps p is true and my knowledge of it is safe, even though there are no exceptionless Carnapian definitions that allow me to "reduce" p to claims about my sensory states. This suggests that genuine knowledge can occur without the possibility of Carnap's "reduction," and so undermines his inference from the fact that we do have genuine knowledge of physical objects based on our sense experience to the conclusion that a reduction of the physical to the psychological *must be possible*. (iii) Finally, it seems possible both that I have

[15] Williamson (2000).

knowledge of the physical and that there is Carnapian "reduction" to the psychological, but only by virtue of an *accidental generalization*. In this case the "successful reduction" sheds no light on the knowledge I actually do have. Taken together, (i)–(iii) undermine both the idea that a Carnapian "reduction" of the physical to the psychological must be possible and the idea that when such a reduction is possible, it can be used to explain our knowledge.

Similar conclusions can be drawn about Carnap's other imagined reductions, including, most importantly, the supposed possibility of reducing our knowledge of physically fundamental things—e.g., electrons or space-time points—to our knowledge of everyday physical things. That reduction, suggested in the 1961 preface to the second edition of the *Aufbau*, would suffer from defects similar to those of the *Aufbau* reduction of the physical to the autopsychological, *if reduction were still conceived as requiring definitions*, as it was in the *Aufbau*.[16] Carnap's grounds for believing the three physicalistic reductions mentioned in the *Aufbau* to be possible were different from his grounds for believing in the possibility of a reduction of the physical to the psychological. Each imagined physicalistic reduction starts with unobservable physical entities posited by theories thought to provide the best explanation of everyday physical facts we already know. Since the domain of objects to which the reduction aims to reduce everything else is far less securely and extensively known than are the domains of familiar things which are to be reduced, one *can't* argue that reductions *must be possible* because otherwise our knowledge of the reductive base wouldn't provide the justification we know we have for our knowledge of the domains to be reduced. In these cases, our knowledge of the reductive base (such as it is) *doesn't* provide our justification for our nontheoretical knowledge. Rather, our knowledge of the former, such as it is, depends on our knowledge of the latter. Thus, one can't argue that Carnapian reducibility of the familiar to the theoretical *must be possible*, since if it weren't we wouldn't even know the familiar.[17]

Carnap didn't think otherwise. I suspect his justification for the claim that everything must be reducible to the physically fundamental was that the physically fundamental is explanatorily fundamental. He was convinced that all psychological facts supervene on and are explained by physical facts, which in turn supervene on and are explained by the most fundamental physical facts. He also seemed to have been convinced that all things are complicated arrangements of the most fundamental physical

[16] As previously noted, Carnap had by then given up this requirement.

[17] As Carnap (1932/33a) makes clear, he did think that our knowledge of the mental states of others was based on our observations of their behavior, and that such observations could, in principle, provide the basis for definitions of the mental in terms of the physical. However, his reasons for thinking that reductions to the most fundamental elements of physics must be possible went beyond this.

things, and all properties of things are physical properties of varying degrees of complexity. The imagined priority in these physical reductions isn't evidential or epistemic; it is explanatory, and hence, covertly, counterfactual.

Of course, no one has produced a successful reduction of all known psychological claims to physical claims, or of all known physical claims to claims about fundamental physical objects. Still, we might wonder whether we have reason to believe that such a reduction *must* be possible. Without a demonstration that facts of type A can't *explain* facts of type B unless the things of type B are "definable" in terms of the primitive properties and relations applying to things of type A, I don't see that we do. We may also wonder whether, *if such a reduction were possible*, it would serve a theoretically important purpose. Perhaps a reduction of some sort would tell us something important. But that doesn't mean that what Carnap called a reduction in the *Aufbau* would do so. As we have seen, the role in Carnapian reductions of true, though not necessarily known or counterfactual-supporting, universally quantified biconditionals as "definitions," suggests that it wouldn't, because it would be possible for such a "reduction" of B-facts to A-facts to connect B-facts with A-facts that don't explain them.

5. THE AUTOPSYCHOLOGICAL REDUCTION

5.1. The Intolerable Burden of the Autopsychological Reduction

Although Carnap believed that reductions of the psychological to the physical and of the physical to the psychological were equally possible, he gave the autopsychological reduction of the physical, and the general psychological to the individually psychological, pride of place. The reason for this was its presumed epistemic primacy as the basis of all knowledge. Carnap explains his notion of *epistemic primacy* in section 54. In describing the autopsychological reduction, he says:

> The system form which we want to give to our outline of the constructional system is characterized by the fact that it not only attempts to exhibit, as in any system form, the order of the objects relative to their reducibility, but that it also attempts to show their order relative to *epistemic primacy.* An object (or an object type) is called *epistemically primary* relative to another one, which we call *epistemically secondary*, if the second one is recognized through the mediation of the first and presupposes, for its recognition, the recognition of the first.[18]

Carnap applies this notion of *epistemic primacy* to the objects countenanced in his constructional systems in section 58.

[18] Carnap (1928 [1967]), pp. 88–89.

> [T]he recognition of our own psychological processes does not need to be mediated through the recognition of physical objects, but takes place directly. Thus, in order to arrange psychological and physical objects in the constructional system according to their epistemic relation, we have to split the domain of psychological objects into two parts: we separate the *heteropsychological objects* from the *autopsychological objects*. The autopsychological objects are epistemically primary to the physical objects, while the heteropsychological objects are secondary. Thus the sequence with respect to the four most important object domains is: the autopsychological, the physical, the heteropsychological, and the cultural.[19]

The *recognition* that Carnap speaks of in these passages is *cognition* in a broad sense that includes recognizing, or knowing, that an entity has a certain property. When recognizing that one object is so-and-so requires the agent to recognize another object is such-and-such, the former object is epistemically prior to the latter.[20] Thus, he seems to embrace (i) and (ii).

(i) An agent's cognition, and knowledge, of physical objects presupposes the agent's cognition, and knowledge, of the agent's private sensory experiences. Hence, an agent's knowledge of physical objects presupposes knowledge of the agent's sensory experiences.
(ii) An agent's cognition, and knowledge, of the agent's private sensory experiences is direct and unmediated, and so does not presuppose cognition, or knowledge, of physical objects. Hence, an agent's knowledge of the agent's own sensory experience does not presuppose knowledge of physical objects.

The task of the autopsychological reduction is to show how it is theoretically possible for an agent to use knowledge of the phenomenal properties of the agent's sensory experience to derive knowledge of the properties of physical objects in the agent's environment, of other physical objects and other agents, and, ultimately, of whatever can be studied scientifically. This, I take it, was the promise enunciated at the beginning of the *Aufbau,* in section 2.

> Even though the subjective origin of all knowledge lies in the contents of experience and their connections, it is still possible, as the constructional system will show, to advance to an intersubjective, objective world, which can be conceptually comprehended and which is identical for all believers.[21]

The starting points for Carnap's ambitious reduction are not discrete experiences of one or another phenomenal property, e.g., experiences of a red sense datum. Instead, they are fleeting sensory *gestalts* called

[19] Ibid., p. 94.
[20] See Friedman (1992) at pp. 120–21 of Friedman (1999).
[21] Carnap (1928 [1967]), p. 7.

elementary experiences, which include within them everything momentarily seen, heard, touched, tasted, or smelled, bound together in a perceptual whole.[22] As he puts it,

> The elementary experiences are to be the basic elements of our constructional system. From this basis we wish to construct all other objects of prescientific and scientific knowledge, and hence also those objects which one generally calls the constituents of experience or components of psychological events and which are found as the result of psychological analysis (for example, partial sensations in a compound perception, different simultaneous perceptions of different senses, quality and intensity components of a sensation, etc.).[23]

Distinct, undifferentiated elementary experiences are said to be related by a primitive relation of *remembered similarity* (section 78), which is used to generate (i) quality classes (section 81)—e.g., of experiences each of which involves (as a part) seeing a colored spot in a certain part of the visual field—(ii) *sense classes* corresponding to the different sensory modalities including classes containing all and only those with visual experiences (as parts), those with auditory experiences (as parts), etc. (section 85), and (iii) classes corresponding to different phenomenal qualities, including those involving color sensations based on hue, brightness, saturation, and location in the visual field. (Since this last quality, location in the flat visual field, actually involves two dimensions, Carnap defines the *visual sense* as the sense class members of which consist exclusively of experiences the qualities of which have five dimensions [sections 80 and 86].) Finally, an intersubjective public space is supposed to be constructed—a space consisting of different *points* at which various properties including color properties/sensations are "located." Eventually, the construction is supposed to include physical objects and other agents, with their own experiences. For Carnap, the crucial requirement is that the construction must yield propositional contents that can be apprehended, believed, and known by all. Somehow these *objective* contents must be abstracted from the *subjective* contents of different individuals. The challenge was to explain how this can be done by "defining" all concepts needed to reconstruct our *common* knowledge from primitive properties of the *private*, undifferentiated sensory inputs of each individual.

Carnap articulates the burden of meeting this challenge in section 66 of the *Aufbau*.

> If the basis of this construction is autopsychological, then the danger of subjectivism seems to arise. Thus, we are confronted with the problem of how we can achieve objectivity of knowledge with such a system form. The requirement that knowledge be objective can be understood in two senses. It could mean

[22] Ibid., section 67.
[23] Ibid., p. 109.

> objectivity in contrast to arbitrariness: if a judgment is said to reflect knowledge, then this means that it does not depend on my whims. Objectivity in this sense can obviously be required and achieved even if the basis for knowledge is autopsychological. Secondly, by objectivity is sometimes meant independence from the judging subject. It is precisely the intersubjectivity which is an essential feature of "reality"; it serves to distinguish reality from dream and deception. Thus, especially for scientific knowledge, intersubjectivity is one of the most important requirements. *Our problem is how science can arrive at intersubjectively valid assertions if all its objects are to be constructed from the standpoint of the individual subject, that is, if in the final analysis all statements of science have as their object only relations between "my" experiences. Since the stream of experience is different for each person, how can there be even one statement of science which is objective in this sense (i.e., which holds for every individual, even though he starts from his own individual stream of experience)?* The solution to this problem lies in the fact that, even though the *material* of the individual streams of experience is completely different, or rather altogether *incomparable*, since a comparison of two sensations or two feelings of different subjects, so far as their immediately given qualities are concerned, is absurd, certain *structural properties* are analogous for all streams of experience. Now if science is to be objective, then it must restrict itself to statements about such structural properties.[24]

The problem is starkly put. The phenomenal content of my sensory experience is private to me. To take a simple example, suppose I have a visual experience which I describe to myself as that of "a circular red dot against a white background." Imagine that, in speaking to myself thusly, I use the words 'red', 'white', and 'circular' to designate phenomenal properties of my experience. Carnap seems to suggest that the proposition I express when whispering (1) under my breath is something I could know to be true, even though that knowledge couldn't be shared by anyone else, and so would be purely subjective.

1. I am seeing (visualizing) a circular red dot against a white background.

In what sense couldn't that purported knowledge be shared? Well, assuming that no one can know the phenomenal properties of my sense data (even if I try to tell them), no one else can know that I, Scott Soames, am having an experience with the phenomenal content reported. What about that proposition I use (2) to express?

2. Someone is seeing (visualizing) a circular red dot against a white background.

Obviously, I could both know that proposition and use sentence (2) to express it. Could anyone else? They could, if (a) like me, they use words like

[24] Ibid., pp. 106–7, my emphasis.

'red', 'white', and 'circular' to designate phenomenal properties of their visual experience and (b) their visual experiences have the same phenomenal properties as mine. Nothing we have said so far rules out different agents knowing the same proposition involving phenomenal properties of conceptually private experiences; nor is it ruled out that they share the *belief* that they both know it. What is ruled out is that they *know* that they both know it.

Might Carnap have something stronger in mind? Well, he does say that "the *material* of the individual streams of experience is completely different, or rather altogether *incomparable*, since a comparison of two sensations or two feelings of different subjects, so far as their immediately given qualities are concerned, is *absurd*." If comparing the phenomenal qualities of private experiences of different subjects is absurd, perhaps *the claim that these qualities are the same* for two subjects is also absurd. Suppose it is. We then get both the result that no two agents know any single proposition about a phenomenal property of private experiences and the result that such common knowledge is impossible. Why might one take the claim that there is such common knowledge to be absurd or impossible? Perhaps because one thinks the supposition of such common knowledge is *meaningless*. But then one can say more. If it is meaningless to claim that the phenomenal properties of private visual experiences of Agent 1 are the same as the phenomenal properties of such experiences of Agent 2, then, surely, (3a) is meaningless, in which case *the pair* of claims (3b) and (3c) is too. But, then, if one of the two must be meaningless, it seems plausible to suppose that both are.

3a. P is a phenomenal property of some private visual experiences of A1 and P is also a phenomenal property of some private visual experiences of A2.
b. P is a phenomenal property of some private visual experiences of A1.
c. P is a phenomenal property of some private visual experiences of A2.

This is tantamount to the claim that there are no phenomenal properties, and hence no knowledge, whether shared or not, of propositions involving such properties. Although I don't think Carnap accepted that conclusion in the *Aufbau*, it is unclear how he would have blocked it.

Nevertheless, it is (relatively) clear how he proposed to solve the problem of achieving objective—i.e., sharable and known to be sharable—knowledge. He must, he thought, eliminate subjective content from what is known by abstracting away from all "material content" so as to arrive at knowledge of *purely structural propositions*. He announces that goal in section 16.

> [E]ach scientific statement can in principle be transformed into a statement which contains only structural properties and the indication of one or more object domains. Now, the fundamental thesis of construction theory . . . asserts that fundamentally there is only one object domain and that each scientific

> statement is about the objects in this domain. Thus it becomes unnecessary to indicate for each statement the object domain, and the result is that *each scientific statement can in principle be so transformed that it is nothing but a structure statement*. But this transformation is not only possible, it is imperative. *For science wants to speak about what is objective, and whatever does not belong to structure but to the material (i.e., anything that can be pointed out in a concrete ostensive definition) is, in this analysis, subjective*. . . . From the point of view of construction theory, this state of affairs is to be described in the following way. The series of experiences is different for each subject. If we want to achieve, in spite of this, agreement in the names for the entities which are constructed on the basis of these experiences, then this cannot be done by reference to the completely divergent content, but only through the formal description of the structure of these entities. However, it is still a problem how, through the application of uniform construction rules, entities result which have a structure which is the same for all subjects, even though they are based on such immensely different series of experiences. This is the problem of inter-subjective reality.[25]

Achieving intersubjective objectivity is the burden of the autopsychological reduction. The burden is unbearable because the Carnapian conditions imposed on solving the problem are unsatisfiable. It will not do to replace one-place phenomenal properties with n-place phenomenal relations—as if that would render the propositions "structural," and hence objective, in the required sense. Rather, *all phenomenal properties and relations* must, somehow, be defined away. But that is impossible. Since this is the *autopsychological* reduction, the only properties and relations—apart from purely logical properties and relations—that remain after the reduction of the physical and the general psychological to the autopsychological are properties and relations applying exclusively and transparently to private experiences of an individual agent. For Carnap, there can be no *objective* (sharable and known to be sharable) knowledge of these.

5.2 Carnap's Unsuccessful Attempt to Secure Objectivity

Although the problem appears to be elementary, the complexity of Carnap's envisioned constructions—involving successively greater abstractions from one's undifferentiated private experiences—obscures the difficulty by all but hiding it under a mass of complicated detail. Earlier I mentioned his definition of the visual sense in section 86 as the *sense class* members of which include experiences the qualities of which involve five dimensions. This may seem to give a purely *structural* characterization of a concept that applies to all agents equally—and hence to be a proper subject of *objective knowledge*.

[25] Ibid., p. 29, my emphasis.

This is an illusion. We may grant that parallel definitions of the *visual sense* can be given for all normally sighted subjects. For the definition to work in any given case, the visual experiences of the agent must include those the qualities of which involve three different dimensions—hue, saturation, and brightness (conceived as features of properties of private sensory experiences)—along with the two dimensions required for location in the flat visual field. But if, as Carnap insists, it is absurd to compare the phenomenal *red* of my experience with that of yours, then it is no less absurd to compare the phenomenal hue, saturation, or brightness of an aspect of one of my experiences with those of yours. So, if the absurdity of the former makes propositions about phenomenal red incapable of being objectively known, then the absurdity of the latter must make propositions about phenomenal hue, saturation, and brightness incapable of being objectively known.

Finally, if none of these are possible objects of genuinely *objective knowledge,* then the objectivity of the concept *visual experience* as defined by Carnap must be suspect. There is nothing magic about the number five. We have no reason to think it is impossible for an agent with *no* visual experiences to have *other* perceptual experiences involving qualities with exactly five dimensions. It is true that Carnap does not require his "definitions" to be necessary truths, and so is indifferent to the observation that possible agents might have nonvisual experiences with exactly five dimensions.[26] But his reply misses two points. First, part of what we *know* is that we have *visual experiences,* as opposed to simply having experiences involving qualities with five dimensions. That too should be intersubjectively available objective knowledge, which ought to, but cannot, be captured by the autopsychological reduction. Second, Carnap's definition of the *dimensions of a sense class* makes use of the primitive two-place relation on private experiences of *recollected similarity*. But just as I cannot compare my experienced phenomenal colors with those of other agents, so I cannot compare my *recollected similarity* relation on my experiences with corresponding relations on the experiences of others. Since the notion *the dimensions of a sense class* is, for Carnap, definable using *recollected similarity*, I can no more compare the number of dimensions inherent in qualities of my visual experience (Carnap's *quantifiable structure*) with the number of those inherent in my neighbor's experience, than I can compare Carnapian *material qualities* of the two streams of private experiences. Hence, it appears that his strategy of using structure to secure objectivity was bound to fail.

It was bound to fail, if Carnap's *purely structural statements* themselves presupposed the primitive relation *recollected similarity* of the autopsychological reduction. Surprisingly, Carnap recognized this. Thus, in section 153, he proposes eliminating even that dependence.

[26] See ibid., p. 140.

> Every constructional system rests upon basic relations which are introduced as undefined basic concepts. Thus all constructed objects are complexes of the basic relations. *All statements that occur in the constructional system are statements about nothing but the basic relation.* . . . However, this characteristic of the statements of a constructional system is not in harmony with the earlier thesis that statements of science must be purely structural. . . . A purely structural statement must contain only *logical symbols*; in it must occur no undefined basic concepts from any empirical domain. Thus, after the constructional system has carried the formalization of scientific statements to the point where they are merely statements about a few . . . [or, in the case of the autopsychological reduction only one] basic relations the problem arises whether it is possible to complete the formalization by *eliminating from the statements of science those basic relations* as the last nonlogical objects.[27]

Surely this is incoherent. If the resulting statements of the constructional system are purely logical, they have no empirical content. Scientific knowledge will not have been rendered objective but obliterated.

Nothing in the *Aufbau* is more stunning than Carnap's failure to recognize this. Part of the reason for his failure may have been the dizzying abstraction with which he pursued the project. Even so, it is not easy to explain how he overlooked the fundamental point. The crucial sections of the *Aufbau* in which he pulls the wool over his own eyes are 153–55. The best summary of this material that I know of is given by Michael Friedman. It begins as follows.

> How is it possible to eliminate even the primitive nonlogical concepts from a constructional system? The method that suggests itself to Carnap is again the method of purely structural definite description. In constructing other objects from our nonlogical primitive(s), we will make essential use of certain empirical facts. In Carnap's [autopsychological] system, for example, we make essential use of the (putative) fact that there is one and only one sense modality based on *Rs* [*recollected similarity*] that is exactly five-dimensional. . . . We could define *Rs*, for example, as the unique basic relation such that there is one and only one sense modality based on it having exactly five dimensions. . . . But a final difficulty now arises. . . . [T]he existence claim implicit in our definition of the basic relation(s) [of *recollected similarity*] will be a logico-mathematical truth [it will be a logical truth that there is at least one abstract relation R such that something with exactly five formal features of a certain structural sort is definable from R], and the uniqueness claim [that there is only one such R] will, in general, be a logico-mathematical falsehood.[28]

As Friedman points out, Carnap notices this problem and attempts a fix.

[27] Ibid., pp. 234–35.
[28] Friedman, Michael (1987) at pp. 102–3 of the reprinting in Friedman (1999).

> Carnap responds then precisely by restricting the range of our variable [over relations]: we are not to consider all relations—which, as mere mathematical sets of pairs, may be "arbitrary unconnected pair lists"—but we are to restrict ourselves to "experienceable, 'natural' relations" [Carnap's words], or what Carnap calls "*founded*" relations (section 154). Carnap next makes the extraordinary suggestion that this notion of *foundedness* may itself be considered a basic concept of logic (section 154), and he completes the "elimination of the basic relation" thusly (section 155): *Rs* is the unique *founded* relation satisfying the chosen empirical conditions (section 155)![29]

This is no fix. Either (i) Carnap has traded one supposedly objectivity-blocking autopsychological primitive relation applying to private experiences for another, or (ii) he has destroyed the autopsychological reduction by introducing an empirical primitive it cannot accommodate, or (iii) he has employed a genuine concept of logic, in which case he has drained his unification of science of all empirical content.

Carnap's failure was not due to lack of ingenuity. The basic problem he set for himself is unsolvable—namely, to explain how it is possible for our sharable, and known to be sharable, common knowledge of an intersubjectively available world to arise from a purely subjective starting point. The problem is unsolvable because the fundamental idea driving the autopsychological reduction is false. Our real starting point is not purely subjective. We do *not* cognize physical objects by cognizing private sensory experience. Although empirical knowledge requires one to *have* sensory experiences, it doesn't require one to *cognize* one's experiences (or any purely private entities they may involve). One doesn't have to perceive the epistemically private, to think about the epistemically private, to predicate properties of it, or to know truths about it in order to have beliefs about, and knowledge of, the intersubjectively available world. When this mistake is eliminated, one is *not* driven to the incredible conclusion that objective—sharable and known to be sharable—knowledge of the world requires the propositions we know to be true to be purely structural.

On the contrary, if one gives up the autopsychological reduction in favor of a physicalistic reduction, the propositions that can be objectively known by different people can include familiar, nonstructural, intersubjectively available, physical-object contents. In short, Carnap's problem arose from his phenomenalism. He wasn't an epistemic foundationalist who was driven to the phenomenal by the need for empirical certainties. But he was a psychological phenomenalist whose methodologically solipsistic starting point generated a pseudo-problem involving objectivity, to which his structuralist thesis appeared as the only possible pseudo-solution.

[29] Ibid., pp. 103.

5.3. The Flawed Treatment of Self and Others

Carnap's confusions concerning objectivity are the most glaring problems for the autopsychological reduction, but they aren't the only ones. The reduction also founders on a flawed account of *the self* and its experiences. To the extent that it makes sense for me to talk of "my self" at all, what the expression picks out is not any part of me, or any entity distinct from but related to me, but just me. At any rate, "myself" does that, and it's not clear what "my self" does, if it doesn't do that. Nevertheless, the latter expression is typically used in philosophy when discussing only a small range of facts about myself—facts private to me about what I am experiencing, which I know in a way I don't know anything else, and which no one else knows about me in that way. For reasons like this, *the self* is often conceived as the one that thinks and experiences in this private way. What is this experiencer? Some say it is a *Cartesian substance,* some it is a *Kantian unity of apperception,* and some say it is a *Humean collection of experiences.* It is a virtue of the *Aufbau* that it doesn't say any of these things. But it isn't easy to pin down what exactly it does say.

In section 64, Carnap calls the autopsychological reduction *solipsistic*, because its base elements are the private experiences of a single agent. Nevertheless, he assures us (i) that the resulting construction doesn't say that there is only a single agent, and (ii) that the experiences that constitute *the given*—which are the basis of the reduction—don't presuppose the existence of any agent at all.

> The autopsychological basis is also called *solipsistic*. We do not thereby subscribe to the solipsistic view that only one subject and its experiences are real, while other subjects are nonreal. The differentiation between real and nonreal objects does not stand at the beginning of a constructional system. As far as the basis is concerned, we do not make a distinction between experiences which subsequent constructions [above the lowest level] allow us to differentiate into perceptions, hallucinations, dreams, etc. . . . The basis could also be described as *the given,* but we must realize that this does not presuppose somebody or something to whom the given is given.[30]

> The expressions "autopsychological basis" and "methodological solipsism" are not to be interpreted as if we wanted to separate, to begin with, the "*ipse*", or the "self", from the other subjects, or as if we wanted to single out one of the empirical subjects and declare it to be the epistemological subject. At the outset [i.e., at the base level of the reduction], we can speak neither of other subjects nor of the self. Both of them are constructed simultaneously at a higher level. . . . In our system form [the autopsychological reduction] the basic elements are to be called experiences of the self *after* the construction

[30] Carnap (1928 [1967]), pp. 101–2.

> has been carried out. . . . [T]he characterizations of the basic elements . . . as "autopsychological", i.e. as "psychological" and as "mine", becomes meaningful only after the domains of the nonpsychological (to begin with, the physical) and of the "you" have been constructed . . . Before the formation of the system, the *basis is neutral* in any system form; that is, in itself, it is neither psychological nor physical."[31]

> *Egocentricity is not an original property of the basic elements* of the given [i.e., they are not so characterized at the lowest level]. To say that an experience is egocentric does not make sense until we speak of the experiences of others which are constructed from "my" experiences. We must even deny the presence of any kind of duality in the basic experience, as it is often assumed (for example, as "correlation between object and subject" or otherwise).[32]

In these passages, we are told that the base elements of the autopsychological reduction include experiences but no experiencers. This doesn't mean that those experiences are *not* experiences of a single agent; in fact, they are so characterized by Carnap at higher levels of the reduction. It does mean that the experiences—out of which all other things, including other agents, are "defined"—are conceptually prior to the thinker or experiencer who has them. This, I believe, is incoherent. Just as it is incoherent to suppose one could conceive of an activity like running without thereby conceiving a physical agent capable of running, so it is incoherent to suppose one could conceive of an activity like perceiving or thinking without thereby conceiving a cognitive agent who perceives or thinks. The key Carnapian primitive in constructing the required definitions is *recollected similarity,* which applies to pairs of experiences and is used to group them into classes. What is it for *experience 1* to bear this relation to *experience 2*? Carnap tells us in section 78: it is for experience 1, which occurred in the past, *to be remembered as similar to* experience 2, which currently occurs. To be remembered by whom? Carnap's characterization presupposes someone, some agent A, who *remembers* having experience 1 and finds it similar to experience 2. After all, individual *experiences*—which are the only elements at the base level of the autopsychological reduction—don't *remember* anything, nor do pairs of them get together and come to the shared conclusion that they are similar.[33] Since *no agents* are recognized at this level, Carnap's relation *recollected similarity* is incoherent. Hence, the autopsychological reduction can't get off the ground.

A different problem arises when we consider not simply the base level of the reduction, but the imagined unification of science that is supposed to be achieved by the reduction as a whole. Remember, the unification of

[31] Ibid., pp. 103–4.
[32] Ibid., pp. 104–5.
[33] On p. 127 Carnap says that "recollected similarity holds between x and y" means "x and y are recognized as part similar through the comparison of a memory image of x with y."

science resulting from "reducing" all claims about the physical and the heteropsychological to the autopsychological is supposed to be noncompetitive with, and representationally identical to, purely physical reductions like the one imagined in the following passage, cited earlier.

> If it is not required that the order of construction reflect the epistemic order of objects, other systems are also possible. . . . Since all cultural objects are reducible to psychological, and all psychological to physical objects, the basis of the system can be placed within the domain of physical objects. Such a system form could be called *materialistic*. . . . However, it is important to separate clearly the logico-constructional aspect of the theory from its metaphysical aspect. From the logical viewpoint of construction theory, no objection can be made against scientific materialism. Its claim, namely, that all psychological (and other) objects are reducible to physical objects is justified.[34]

The physicalistic reduction imagined here contrasts with the autopsychological reduction. Although both envision an exceptionless correlation of mental events or states (e.g., thoughts, perceptions, and other experiences) with physical events or states (e.g., neurological events or states), in the physicalistic reduction the former are "defined" in terms of, and hence "reduced" to, the latter, while in the autopsychological the direction is reversed.[35]

To simplify, the physicalistic reduction allows us to truly say that all sensations are nothing but brain states, while the autopsychological reduction allows us to say that all brain states are nothing but sensations. Carnap's simultaneous embrace of these claims stems from his view that the unifications of science resulting from the two reductions represent the world as being in precisely the same state. In section 3 of this chapter, I suggested that the best explanation for this is one that reconstructs his position as adopting a version of holistic verificationism. On this view, the content of an individual claim—e.g., that all sensations are brain processes or that all brain processes are sensations—is, roughly, that which it contributes to the content of the overall theory (in this case to the unification of science) of which it is a part. The two claims are compatible, and even complementary, if (i) the two unifications of science make the same observational predictions and (ii) the two claims make comparable contributions to the two unified theories of which they are parts.

Now back to the self. Imagine I wake up in the dark unable to move, after being drugged. My only sensations are of a tiny point of light and a faint sound of music. Although I am able to think perfectly well, I am utterly in doubt about what has happened. In such a pseudo-Cartesian situation I might know little else than that since *I have thoughts and experiences, I must exist*. What is it that I know? Certainly not simply that

[34] Ibid., p. 95.
[35] The possible observations establishing this correlation are discussed in section 168.

there are thoughts and experiences, or even thoughts and experiences of a certain type. That could be true even if the propositions I, in fact, know were false. For the same reason, what I know is not simply that someone is having thoughts and experiences.[36] Suppose further, with Carnap, that materialism is correct and that, like every other human being, I am nothing more than a certain complex physical system. Then, in knowing that I exist, *I know of a certain human being, which is nothing more than a physical system, that it exists.* Still, I may not know that anything human or even physical exists. Moreover, what I know is different from what you would know, if you were in an identical situation.

How, in light of this, could Carnap's autopsychological theory of the world possibly capture my knowledge of my own existence and sensations? It could do so only if (i) it were capable of specifying what uniquely distinguishes me from all other agents and (ii) that information were extractable from the contributions my knowledge of myself makes to the observational predictions of the total theory. Since Carnap's autopsychological reduction doesn't satisfy these conditions, he cannot capture the most elementary knowledge individuals have of themselves.

6. THE SCOPE OF CARNAPIAN TRUTH, KNOWLEDGE, AND SCIENCE

The autopsychological reduction was, for the reasons indicated, a disaster. To salvage something from it, one must eliminate both private experiences as items knowledge of which ground all other knowledge and *definitional reduction* of higher to lower domains as the form of a system of unified science. Doing both has allowed more recent philosophers to focus on specifying what scientific theories are, what their intersubjective observational evidence consists in, and what, if anything, beyond equivalence of observational predictions is required in order for different theories to represent the world as being in the same state. It has also allowed philosophers to pose answers to sophisticated questions far beyond those

[36] Carnap appears to be oblivious to these obvious points. Thus, on p. 261 (ibid.) of the section "The Problem of the Self" he says, "The existence of the self is not an originally given fact. The *sum* does not follow from the *cogito*; it does not follow from 'I experience' that 'I am,' but only that an experience is. The self does not belong to the expression of the basic experience at all, but is constructed only later. . . . Thus a more fitting expression than 'I experience' would be 'experience' or still better 'this experience.' Thus, we ought to replace the Cartesian dictum by 'this experience: therefore this experience is,' and this is of course a mere tautology." It is not, of course, a tautology, since it is not even a well-formed sentence. The whole passage is a combination of nonsense and falsehood. One can't replace a sentence "I experience" with a noun "experience" or a noun phrase "this experience." The famous critic of Heidegger's "Nothing nothings" seems in his early years to have been no slouch in the production of nonsense himself.

envisioned in the *Aufbau* concerning our justification for accepting scientific theories, as well as for believing, or knowing, them to be true. The abandonment of *definitional reduction* as the means by which theoretical claims *must* be related to evidence has also reduced the motivation for supposing that there *must* be a way of unifying all of science into a single hierarchically interconnected system. Finally, the recognition that much of one's knowledge—e.g., my knowledge *that I exist, that I am now having various experiences,* and *that I am not Saul Kripke*—is irreducibly singular while still being fully objective has made it less plausible to expect genuinely scientific knowledge to encompass all objective knowledge.

These limitations are foreign to Carnap. At the end of the *Aufbau*, in sections 179 and 180, he articulates his vaulting conception of the aims of science and the scope of scientific knowledge.

> The aim of science consists in finding and ordering the true statements about the objects of cognition (not all true statements . . . ; we do not undertake to discuss the teleological problem . . . at this point).[37]

Here it is suggested that with one possible exception—teleological truths—the task of science is to discover all truths about "objects of cognition." Since those are presumably *things we can think about,* it sounds like the domain of science includes all truths (except teleological truths, if there are such). This impression is reinforced two pages later, when Carnap characterizes science as "the system of conceptual knowledge."

> *Science, the system of conceptual knowledge, has no limits.* But this does not mean that there is nothing outside of science. . . . The total range of life has still many other dimensions outside of science, but within its dimension, science meets no barrier. . . . When we say that scientific knowledge is not limited, we mean: *there is no question whose answer is in principle unattainable in science.* . . . It is occasionally said that the answer to some questions cannot be conceptualized; that it cannot be formulated. But in such a case, the question itself could not have been formulated.[38]

Here we learn that that every question that can be scientifically formulated can be answered. Two issues remain: *What is it for a scientific question to be answered?* and *Are there genuine nonscientific questions that might nevertheless have true, and even knowable, answers?*

Carnap addresses the first of these as follows.

> Now, if it is the case that a genuine question is posed, what are the possibilities of giving an answer? In such a case, a statement is given; it is expressed through conceptual symbols in formally permissible combination. Now, in principle, every legitimate concept of science has a definite place in the

[37] Carnap (1928 [1967]), p. 288.
[38] Ibid., p. 290.

> constructional system. . . . *We now replace the sign for each of these concepts as it occurs in the given sentence by the expression which defines it in its constructional definition*, and we carry out, step by step, further substitutions of constructional definitions. We already know that, eventually, the sentence will have a form in which . . . it contains only signs for basic relations [recollected similarity in the autopsychological reduction]. . . . In keeping with the tenets of construction theory, *we presuppose that it is in principle possible to recognize whether or not a given basic relation holds between two given elementary experiences*. Now, the state of affairs in question is composed of nothing but such individual relation extension statements [about recollected similarity], where the number of elements [private experiences] which are connected through the basic relation [recollected similarity] . . . is finite. *From this it follows that it is in principle possible to ascertain in a finite number of steps whether or not the state of affairs in question obtains and hence that the posed question can be answered.*[39]

Here Carnap presupposes what I have already argued should be rejected—definitional reducibility to the subjective experiences that constitute the base elements of the autopsychological construction. So, if my arguments are well taken, his conclusion should not be ours. Nevertheless, his claim is worth noting. *All scientific questions can be answered, because all meaningful scientific statements are, in principle, conclusively verifiable, and hence capable of being known to be true, or false*. We are here approaching the signature claim of logical empiricism. The only remaining issue is whether there are genuinely meaningful nonscientific questions the answers to which can be verified and hence known.

Carnap immediately takes up this issue in section 181.

> According to the above-indicated position, *conceptual knowledge* does not meet any limitations in its own field; nevertheless, it is an open question whether it is perhaps possible to gain insights in a manner which lies outside *conceptual knowledge* and which is inaccessible to *conceptual thinking*. . . . Unquestionably, there are phenomena of faith, religious and otherwise, and of intuition; they play an important role, not only for practical life, but also for cognition. Moreover, it can be admitted that, in these phenomena, somehow something is "grasped," but this figurative expression should *not lead to the assumption that knowledge is gained* through these phenomena. What is gained is a certain *attitude*, a certain psychological state, which, under certain circumstances, can indeed be favorable for obtaining certain insights. *Knowledge, however, can be present only when we designate and formulate, when a statement is rendered in words or other signs.* Admittedly the above-mentioned states put us occasionally in a position of asserting a statement or ascertaining it truth. *But it is only this articulable, hence conceptual, ascertainment which is knowledge*; it must be carefully distinguished from that state itself.[40]

[39] Ibid., pp. 291–92, my emphasis.
[40] Ibid., pp. 292–93, my emphasis.

We have already been told that all conceptual knowledge falls within the domain of science. It is here suggested that (i) there is no knowledge outside that domain, and (ii) what falls outside that domain isn't stateable in words or symbols. In the next paragraph Carnap characterizes the nonconceptual deliverances of faith or intuition as *ineffable*, paraphrasing the *Tractatus*: "For, we cannot speak of question and answer if we are concerned with the ineffable."[41]

All of this suggests that for Carnap, at the very end of the *Aufbau,* the domain of science encompasses all knowledge and all truths. Since no stateable question or statement falls outside that domain, every truth-apt—i.e., cognitively meaningful—sentence is either conclusively verifiable or conclusively falsifiable, and hence capable of being known to be true or known to be false. This is classical logical empiricism of the sort espoused at about the same time by Schlick, under the influence of the *Tractatus.*

7. THE LEGACY OF THE *AUFBAU*

Viewed from today's perspective, 87 years after the publication of the *Aufbau,* that work is apt to seem more Kantian than contemporary. Like the *Critique of Pure Reason*, it purported to set out an encompassing framework within which all human knowledge can be explained and beyond which human knowledge is impossible. Unlike Kant's system, which was a grandly schematic piece of aprioristic philosophical psychology through which the science of his day was to be understood and the limits of human reason were to be set, Carnap's system was a grandly schematic piece of philosophical logic and linguistic analysis in which the vastly more complex science of his day was to be explicated, and the limits of meaningful thought and talk were to be delineated. However, this impression of old wine in new bottles should not be pushed too far. The stunning advances of the late nineteenth and early twentieth centuries in physics, logic, and mathematics raised a host of new philosophical questions requiring new philosophical approaches. Carnap, Schlick, and their fellow logical empiricists understood these advances, took them seriously, and struggled to make science itself, and its relation to all areas of human thought, central to philosophy in a way it had not been before. The success of the *Aufbau* lies not in its substantive philosophical doctrines, but in the agenda centered on science, logic, and language that it helped to set for philosophy.

[41] Ibid, p. 293. Four pages later Carnap praises Wittgenstein and quotes section 6.5 of the *Tractatus*: "When the answer cannot be put into words, neither can the question be put into words. The riddle does not exist. If a question can be framed at all, it is also possible to answer it."

CHAPTER 7

The Heyday of Logical Empiricism

1. The Vienna Circle in 1930
2. Schlick and Carnap: The Turning Point in Philosophy, the Logical Analysis of Language, and the Elimination of Metaphysics
3. Hans Hahn: The Linguistic Theory of the A Priori
4. Schlick's Foundation of Knowledge
5. Hempel: Truth, Confirmation, and Certainty
6. Reichenbach: The Elimination of Truth

1. THE VIENNA CIRCLE IN 1930

By 1930, Carnap's *Aufbau* (1928) and the logical empiricist manifesto "The Scientific Conception of the World"—Hahn, Carnap, and Neurath (1929)—had been published, the *Tractatus* had been digested, and Wittgenstein's influence on members of the Vienna Circle had been firmly established (between 1926 and 1929). The worldview called for in the manifesto was already in place. In keeping with the document's identification of Albert Einstein, Bertrand Russell, and Ludwig Wittgenstein as "leading representatives of the scientific world-conception," the worldview combined a tractarian conception of language, philosophy, and the limits of intelligibility with a verificationist conception of knowledge and meaning, and the use of the new (Russellian) logic and (Einsteinian) physics as paradigms of a priori and empirical knowledge, respectively. Among the important themes were the collapse of apriority and necessity into analyticity (or "logical truth"), the abolition of religion, metaphysics, and normative theory, the apotheosis of the scrutable, the unification of science (into a single explanatory system), the conception of philosophy as the logical analysis of science, and the dismissal of questions about the relationship between linguistic representations of reality and reality itself.

I have already outlined how Carnap linked several of these themes in the *Aufbau*—including the task of philosophy, the unification of science,

the verificationist reading of the *Tractatus,* and the elimination of nonscientific claims to worldly insight. Recall the opening of section 180:

> *Science, the system of conceptual knowledge, has no limits . . . there is no question whose answer is in principle unattainable in science*.[1]

So, every question can be answered scientifically. What does such an answer amount to?

> [I]f it is the case that a genuine question is posed, what are the possibilities of giving an answer? . . . [E]very legitimate concept of science has a definite place in the constructional system. . . . *We now replace the sign for each of these concepts as it occurs in the given sentence by the expression which defines it in its constructional definition*, and we carry out, step by step, further substitutions of constructional definitions. . . . [E]ventually, the sentence will have a form in which . . . it contains only signs for basic relations [recollected similarity in the autopsychological reduction]. . . . [W]e presuppose that it is in principle possible to recognize whether or not a given basic relation holds between two given elementary experiences. Now, the state of affairs in question is composed of nothing but such individual relation extension statements, where the number of elements which are connected . . . is finite. *From this it follows that it is in principle possible to ascertain in a finite number of steps whether or not the state of affairs in question obtains and hence that the posed question can be answered*.[2]

Carnap concludes that *all scientific questions can be answered, because all meaningful scientific statements are conclusively verifiable or falsifiable, and so capable of being known to be true, or false*. Although he doesn't give a definition of what it is for a statement to be verifiable, he is confident that whenever a statement is meaningful, its truth or falsity can be conclusively be determined.

His conclusion covers all meaningful conceptual claims.

> [C]onceptual knowledge does not meet any limitations in its own field. . . . Knowledge . . . can be present only when we designate and formulate, when a statement is rendered in words or other signs. . . . [I]t is only this articulable, hence conceptual, ascertainment which is knowledge.[3]

We have already seen that all conceptual knowledge falls within the domain of science. It is here suggested that there is no knowledge not stateable in words that falls outside that domain. Next, Carnap calls the nonconceptual deliverances of faith or intuition *ineffable*, paraphrasing the *Tractatus*: "For, we cannot speak of question and answer if we are concerned with the

[1] Carnap (1928 [1967]), p. 290.
[2] Ibid., pp. 291–92, my emphasis.
[3] Ibid., pp. 292–93, my emphasis.

ineffable." This suggests that for Carnap in 1928, every cognitively meaningful sentence is conclusively verifiable or conclusively falsifiable, and so capable of being known to be true or known to be false. This is classical logical empiricism.

2. SCHLICK AND CARNAP: THE TURNING POINT IN PHILOSOPHY, THE LOGICAL ANALYSIS OF LANGUAGE, AND THE ELIMINATION OF METAPHYSICS

In 1930 Rudolf Carnap and Hans Reichenbach founded *Erkenntnis,* which was, in effect, a house journal for logical empiricism.[4] The lead article in the first issue was Moritz Schlick's "The Turning Point in Philosophy." More a triumphal proclamation of a connected set of philosophical theses than an argument for them, it revealed the astonishing confidence of the logical empiricists. Schlick contended that after millennia of little or no progress, characterized by the "chaos of [philosophical] systems" and the "anarchy of philosophical opinions," "we now find ourselves at an altogether decisive turning point in philosophy, and . . . we are objectively justified in considering that an end has come to the fruitless conflict of systems."[5] Crediting Frege and Russell with pioneering work in logic that made the breakthrough possible, Schlick gives pride of place to the *Tractatus* for being the work that "pushed forward to the decisive turning point."[6]

According to Schlick, Wittgenstein's chief contribution was the identification of language—our chief means of representing reality—as the proper subject matter of philosophy. He says,

> Investigations concerning the human "capacity for knowledge" . . . are replaced by considerations regarding the nature of expression, of representation, i.e. concerning every possible "language". . . . Questions regarding the "validity and limits of knowledge" disappear. Everything is knowable which can be expressed, and this is the total subject matter concerning which meaningful questions can be raised. Consequently there are no questions which are in principle unanswerable, no problems which are in principle unsolvable. What have been considered such up to now are not genuine questions, but meaningless sequences of words.[7]

Schlick's imagined breakthrough came from substituting the question "What do we mean?" for the question "What can we know?" Once we

[4] It ran under that name until 1938, when it was renamed *The Journal of Unified Science (Erkenntnis)*, which operated until 1940, when its publication was halted by World War II. It was refounded as *Erkenntnis* in 1975 by Wilhelm Esler, Carl Hempel, and Wolfgang Stegmuller.
[5] Schlick (1930/31 [1959]), p. 54.
[6] Ibid., p. 54.
[7] Ibid., pp. 55–56.

understand that meaning is verification, we see that every meaningful conjecture is capable of being known to be true or known to be false, in which case there will be no unanswerable questions.

> Whenever there is a meaningful problem one can in theory always give the path that leads to its solution. For it becomes evident that giving this path coincides with the indication of its meaning. . . . The act of verification . . . is always of the same sort: it is the occurrence of a definite fact that is confirmed by observation, by means of immediate experience. In this manner, the truth (or falsity) of every statement, of daily life or science, is determined. . . . Every science . . . is a system of cognitions, that is, of true experiential statements. And the totality of sciences, including the statements of daily life, is the system of cognitions.[8]

The system to which Schlick alludes is one in which all meaningful claims about the world are verifiable or falsifiable. Since their investigation is empirical, they fall outside of philosophy. Nor is philosophy devoted to constructing a system of a priori truths. Like other logical empiricists, Schlick thought that all a priori truths are true in virtue of meaning. Although logic and mathematics aim at discovering bodies of such truths, philosophy's task isn't to carve out and systematize any special class of truths; it is to clarify meanings.

> Philosophy is not a system of statements; it is not a science. . . . The great contemporary turning point is characterized by the fact that we see in philosophy not a system of cognitions, but a system of *acts*; philosophy is that activity through which the meaning of statements is revealed or determined. By means of philosophy statements are explained, by means of science they are verified. The latter is concerned with the truth of statements, the former with what they actually mean.[9]

Also appearing in the first issue of *Erkenntnis* was Carnap's "The Old and the New Logic," which, like Schlick's, announced a new era of philosophy as meaning clarification in the service of science.[10] Like Schlick, Carnap also made extravagant claims about what he took to be the revolutionary impact of "the new logic" on philosophy.

> [I]n the new logic . . . lies the point at which the old philosophy is to be removed from its hinges. Before the inexorable judgment of the new logic, all philosophy in the old sense, whether it is encountered with Plato, Thomas Aquinas, Kant, Schelling or Hegel, or whether it constructs a new "metaphysic of Being" or a "philosophy of spirit," proves itself to be not merely materially false, as earlier critics maintained, but logically untenable and therefore meaningless.[11]

[8] Ibid., p.56.
[9] Ibid., p. 56.
[10] Carnap (1930/31 [1959]).
[11] Ibid., p. 154.

Unlike Schlick, Carnap was expansive about the role of Frege, Russell, and others in developing the new logic.

> The most important stimulus for the development of the new logic lay in the need for a critical re-examination of the foundations of mathematics. . . . Mathematics succeeded in defining . . . such important concepts as limit, derivative, and complex number. . . . People were not satisfied with reducing the various concepts of mathematical analysis to the fundamental concept of number; they required that the concept of number should itself be logically clarified. This inquiry into the *logical* foundations of arithmetic with a *logical* analysis of number as its goal . . . gave especially strong impetus to the development of the new logic. Peano, Frege, Whitehead, Russell and Hilbert were led to do their work on logic primarily for this reason.[12]

Two aspects of the discussion of logic in Carnap (1930/31) stand out. The first is his take on Russell's theory of types as a way of avoiding paradox. In discussing the *heterologicality* paradox, he credits the theory of types with blocking the paradoxical result that the property *not applying to itself* applies to itself iff it doesn't apply to itself. Seeing in this a vindication of type theory as a general constraint on intelligibility, he embraces the idea that any sentence that violates it—e.g., by predicating an n^{th}-level property or relation of anything other than entities of a level lower than n—is neither true nor false, but meaningless. Carnap's enthusiasm was, I think, misplaced. Although the theory of types does block the paradox, nothing so elaborate is needed, since, to put it simply, (i) the first-order claim *There is a property p that applies to all and only those properties that don't apply to themselves* is a straightforward contradiction in classical logic, and so doesn't require a special mechanism to block the paradox. It may also be noticed that the claim $\exists R \forall x\ (Rxx \text{ iff } \sim Axx)$ is unparadoxically true in textbook second-order logic no matter how the predicate *Axy* is interpreted, and no matter whether the domain of first-order quantification includes properties and relations or not. More generally, the reading of *Principia Mathematica* on which the type restrictions are, in fact, genuine constraints on the intelligibility of the (substitutional) quantification employed there weakens the ability of the system to provide a basis for higher mathematics and creates problems for fundamental aspects of both Russell's and Carnap's broader philosophical logic, thereby undercutting the general lesson Carnap draws from it.[13]

The second notable aspect of his discussion of "the new logic" is his acceptance of logicism as an established fact. He credits *Principia Mathematica* with showing "*that every mathematical concept can be derived from the fundamental concepts of logic and that every mathematical sentence (insofar*

[12] Ibid., p. 135, my emphasis.
[13] See Soames (2014), chapter 10, section 4, plus the third section of Soames (2015a).

as it is valid in every conceivable domain of any size) can be derived from the fundamental statements of logic."[14] The first of these claims is defensible in so far as *Principia Mathematica* did show how to define every arithmetical concept, and hence every arithmetically definable mathematical concept, using concepts that might be regarded as logical.[15] But the second claim is puzzling. Russell freely admitted that his logicist reduction of (first-order) Peano arithmetic requires his axiom of infinity, which, by 1919, he didn't take to be a "statement of logic," or even to be knowable a priori.[16] If Carnap wasn't here simply making a mistake, perhaps what he meant by *valid in every conceivable domain of any size* was *valid in every conceivable domain of any size consistent with the "logic" in Principia Mathematica*. If so, he was embracing a shortcoming inherent in the type-theoretic reduction.

Two other facts we now know, but Carnap may not have known then, undermine his claim that *Principia Mathematica* shows every arithmetical truth to be *derivable* from logical truths. The first, established in Gödel (1931), was that simply reducing first-order Peano arithmetic to a "logical system" fails to guarantee that the system can derive all arithmetical truths, because infinitely many of those truths are *not* logical consequences of the first-order Peano axioms.[17] The second pertinent fact, which is a corollary of Gödel (1931), is that although *all* arithmetical truths *are* logical consequences of *second-order* Peano arithmetic, there can be no complete proof procedure for second-order logical truth. Thus the expansive claims Carnap makes about the consequences of Russell's new logic for our understanding of the relationship between logic and mathematics were (perhaps understandably) exaggerated.

This pattern of enthusiastic but unsupported claims about logic and mathematics continues when Carnap turns to the *Tractatus* for inspiration.

> On the basis of the new logic, the essential character of logical sentences can be clearly understood. . . . The usual distinction between fundamental and derived sentences is arbitrary. It is immaterial whether a logical sentence is derived from other sentences. *Its validity can be recognized from its form*.[18]

This remark echoes the proper tractarian rejection of a proof-theoretic conception of what it is to be a truth of logic in favor of some more fundamental conception. However, it is unclear what conception Carnap had in mind. The last sentence of the passage, which claims that validity (logical truth) can be *recognized* by (syntactic) form alone, might be seen as suggesting either (i) that every logical truth can be proved to be such on the basis of its form alone (which holds for first-order, but not higher-order,

[14] Carnap (1930/31 [1959]), pp. 140–41.
[15] See Boolos (1994).
[16] Russell (1919), pp.202–3. See also Soames (2014), chapter 10, section 3.4.
[17] See section 2 of the next chapter.
[18] Carnap (1930/31 [1959]), pp. 141–42, my emphasis.

logic), or, (ii) that there is a way of deciding, for absolutely every sentence, whether it is, or is not, logically true simply by examining its form (which fails even for first-order logic). The error of suggesting that (i) is true is understandable; the error of suggesting that (ii) is true is more serious.

The worry that Carnap may have committed it is reinforced when he continues the passage with a laborious truth-table demonstration that *(A or B) or (~A and ~B)* is a tautology, followed by this general conclusion:

> Such a formula, which depends neither on the meanings nor the truth-values of the sentences occurring in it but is necessarily true, whether its constituent sentences are true or false, is called a tautology. *A tautology is true in virtue of its mere form. It can be shown that all the sentences of logic and, hence, according to the view advocated here, all the sentences of mathematics are tautologies.*[19]

Carnap's reader is invited to think that all logical truths and all mathematical truths have the same status as do tautologies of the propositional calculus—the *decision procedure* for which he had just illustrated. Might Carnap have believed in 1930 that there is always a decision procedure for logical and mathematical truth? Perhaps. The *Tractatus*, by which he was heavily influenced, contains a similar suggestion.[20] Moreover, it was not until 1936 that Church, followed shortly by Turing, proved the undecidability of first-order logic.[21] So the idea that decision procedures might be essential to logic hadn't, in 1930, been proved wrong. Still, it should have been recognized even then that there was no guarantee that the vindication of (i) in Gödel (1930) could be extended to higher-order logical truths, let alone that (beyond the propositional calculus) any vindication whatsoever could be given for the stronger claim (ii).[22] Thus, it was imprudent for Carnap to assume, or to allow his readers to assume, that these were established results.

Carnap next turns to a tractarian-inspired discussion of what is stated by a truth of logic or mathematics.

> If a compound sentence is communicated to us, e.g., "It is raining or it is snowing," we learn something about reality. This is so because the sentence excludes certain of the relevant states-of-affairs and leaves the remaining ones open. . . . If, on the other hand, we are told a tautology, no possibility is excluded. . . . Consequently, we learn nothing about reality from the tautology. . . . Tautologies are, therefore, empty. They say nothing; they have, so-to-speak, zero content. However, they need not be trivial on this account. The

[19] Ibid., p. 142, my emphasis.
[20] See the discussion of this issue in chapter 3, section 4 of this volume.
[21] Church (1936a), and Turing (1936/37), along with other important theorems of Gödel, Tarski, and Rosser, are discussed in chapter 8.
[22] Gödel (1930) is the slightly strengthened published version of his 1929 dissertation. It proves the completeness of systems of proof for first-order logic— i.e., the existence of systems of proof capable of proving all and only first-order logical truths.

> above-mentioned tautology is trivial. On the other hand, there are sentences whose tautological character cannot be recognized on first glance.[23]

The doctrine that all logical, mathematical, and indeed all necessary and a priori, truths have zero content is far removed from the logicism of Frege and Russell. It also has nothing to do with "the new logic" and everything to do with the new philosophy of language of the *Tractatus*. Worse, the idea that all of these truths really "say nothing" is strikingly counterintuitive, and so in need of a powerful defense, which Carnap doesn't give. Surely, to say *that first-order Peano arithmetic is incomplete* is to say something different and more informative than to say *that 0 ≠ 1*—which would not be so if to say both *said nothing*. Moreover, if to assert or believe these truths were to assert or believe nothing (about the world, or about symbols, or about anything else), then presumably to assert or believe their negations—i.e., *that first-order Peano arithmetic is complete* and *that 0=1*—would be to assert or believe everything (including each proposition and its negation). Since it is clearly impossible to (simultaneously) assert or believe everything (and its negation), it would follow that no one has ever asserted or believed a logical, mathematical, necessary, or a priori falsehood.[24] It is striking that Carnap feels no need to explain why this isn't a *reductio ad absurdum* of this tractarian view.

A related doctrine, though not as contentious as the one just questioned, identifies apriority with analyticity (truth in virtue of linguistic convention).

> Mathematics, as a branch of logic, is also tautological. In Kantian terminology: The sentences of mathematics are analytic. . . . Empiricism, the view that there is no synthetic *a priori* knowledge, has always found the greatest difficulty in interpreting mathematics. . . . This difficulty is removed by the fact that mathematical sentences are neither empirical nor synthetic *a priori* but analytic.[25]

This, as we shall see in chapter 10, would also prove to be difficult to defend. As for the alleged transformation of philosophy by "the new logic" (and the new philosophy of language), Carnap concludes his discussion with the following summary:

> Every sentence of science must be proved to be meaningful by logical analysis. If it is discovered that the sentence in question is either a tautology or a contradiction . . . the statement belongs to the domain of logic including mathematics. Alternatively the sentence has factual content, i.e., it is neither

[23] Carnap (1930/31 [1959]), pp. 142–43.
[24] These results follow if (i) when P and Q have the same content ⌜A asserts/believes P⌝ is true iff ⌜A asserts/believes Q⌝ is true, and (ii) if ⌜A asserts/believes P&Q⌝ is true, then ⌜A asserts/believes P⌝ and ⌜A asserts/believes Q⌝ are both true.
[25] Carnap (1930/31 [1959]), p. 143.

> tautological nor contradictory; it is then an empirical sentence. It is reducible to the given and can, therefore, be discovered, in principle, to be either true or false. . . . There are no questions which are in principle unanswerable. There is no such thing as speculative philosophy, a system of sentences with a special subject matter on a par with those of the sciences. To pursue philosophy can only be to clarify the concepts and sentences of science by logical analysis.[26]

In 1932, Carnap published another short article in the second volume of *Erkenntnis,* this time emphasizing the negative lesson of the *Aufbau*. The positive lesson was, of course, that philosophy's chief task in clarifying meaning was to reveal the logical and epistemological structure of science, and to systematize it into a unified whole. The negative lesson, explained in Carnap (1932), was that philosophy must remove *metaphysics* and *normative theory,* which are impediments to achieving that goal.

> In the domain of *metaphysics,* including all philosophy of value and normative theory, logical analysis yields the negative result *that the alleged statements in this domain are entirely meaningless.*[27]

With characteristic thoroughness, Carnap sketches the two main ways in which meaningful *pseudo-statements* arise, through meaningless words and through counter-meaningful combinations of individually meaningful words.

He begins by asking "What is the meaning of a word?" which he identifies with the question "What stipulations, explicit or implicit, give words their meanings?" First, he says, one specifies the syntax of the word, which he takes to be revealed by the simplest, "elementary" sentences in which it occurs. His example of elementary sentences containing the word 'stone' are 'This diamond is a stone' and 'this apple is a stone'.[28] Next, the significance of an elementary sentence S containing the word is given by specifying which sentences S is *deducible* from and which sentences are *deducible* from S. By *deducible*, he means *formally deducible*, where for Carnap *formal deducibility*, *logical consequence*, *necessary consequence*, and *a priori consequence* are one and the same.

Although it is common today to distinguish these four notions, it wasn't common in 1932. To say that B is formally deducible from A (in a given system of proof) is to say that there is a formally correct derivation of B from A, i.e., a finite sequence of formulas connecting A to B in which every line following A is either an axiom or the result of a syntactic transformation of earlier lines sanctioned by a rule of inference. Since formal deducibility is always mechanically checkable by examining the symbols on

[26] Ibid., p. 145.
[27] Carnap (1932 [1959]), pp. 60–61.
[28] Ibid., p. 62.

each line, Carnap took it to be a central notion of *logical syntax* (as we do today). But he also identified it with logical consequence, which we do not define syntactically. It wasn't until Tarski (1935, 1936) that it became common to identify the *semantics* of an interpreted formal language L with a definition of *truth in a model* plus an intended *model* consisting of a domain of objects about which L is used to make claims and an assignment to each nonlogical expression of an element, or a set-theoretic construction of elements, from the domain. Within this perspective, *logical consequence* is *semantically* defined; to say that B is a logical consequence of A is to say that B is true in every model in which A is true. Whether or not the relations *logical consequence* and *formal derivability* have the same extensions in L (relative to a given proof procedure) varies with L. For some languages they do; for some they don't. For languages with unrestricted higher-order quantification, the two relations are never coextensive.

Although Carnap would, in time, see this, he had not done so in 1932. Indeed, in Carnap (1934b), logical syntax was, for him, simply logic.

> What linguists call rules of syntax are indeed such formal . . . rules for the formation of propositions [sentences]. We can see, however, clearly that the transformation rules [i.e., rules of inference], which one usually calls logical rules of deduction, have the same formal, that is, syntactical character. . . . One of the most important concepts of logic and thereby of the logic of science is that of logical inference (Folgerung—entailment). . . . The decisive point is: is it . . . possible to formulate the concept "entailment" purely formally? If the transformation [inference] rules of language are set up purely formally, we call a proposition [sentence] an inference (entailment) of other propositions [sentences] if it can be constructed from those propositions [sentences] by the application of transformation [inference] rules. . . . The question, whether a certain proposition is an inference (entailment) of certain other propositions . . . is answered by a *Combinatorial Calculus or Mathematics of Language,* which rests on the transformation [inference] rules of language, that is what we have called the *syntax* of language. Briefly: "entailment" is defined as deducibility according to the transformation [inference] rules; since these rules are formal, "entailment" is also a formal, syntactical concept.[29]

This passage gives the flavor of Carnap's most advanced pre-Tarskian position. Formal derivability is identified not only with logical consequence, but with entailment, which might otherwise be understood as necessary or a priori consequence.

With this in mind, we return to Carnap's discussion of the meaning of a word, which he took to be given by specifying the meanings of elementary sentences containing it. As we have seen, the meanings of these sentences were to be given by identifying the sentences from which they are formally

[29] Carnap (1934b), pp. 10–11.

deducible plus the sentences (formally) deducible from them. Here is how he puts it.

> [F]or an elementary sentence S containing the word an answer must be given to the following question, which can be formulated in various ways:
>
> (1) What sentences is S *deducible* from, and what sentences are deducible from S?
> (2) Under what conditions is S supposed to be true, and under what conditions false?
> (3) How is S to be *verified*?
> (4) What is the *meaning* of S?
>
> (1) is the correct formulation; formulation (2) accords with the phraseology of logic, (3) with the phraseology of theory of knowledge, (4) with philosophy (phenomenology). Wittgenstein has asserted that (2) expresses what philosophers mean by (4): the meaning of a sentence consists in its truth-condition. ((1) is the "metalogical" formulation; it is planned to give elsewhere a detailed exposition of metalogic as the theory of syntax and meaning, i.e. relations of deducibility.)[30]

Remarkably, Carnap regards (1)–(4) as different formulations of *the same question*. They are not. Indeed, his understanding of deducibility as formal deducibility make (1) and (4) about as distant in content as one could possibly imagine. Syntax is not semantics and sentence structure is not meaning. Surely, we want to object, simply knowing how a sentence S is related to other sentences by syntactic transformations tells us next to nothing about what S means. How, one wonders, could Carnap have thought otherwise?

He thought otherwise because he then had no notion of *truth* as a semantic property of sentences that accurately represent things as being as they really are. Following the *Tractatus*, Carnap regarded attempts to state the relationship between language and the world in virtue of which sentences mean what they do (and so have truth conditions) as misleading, and ultimately meaningless. What such pseudo-statements try to state was, for him, more accurately and unproblematically stated by claims about the syntactic relationships between different linguistic forms. This is why he took (1) to be the *correct formulation* of the question that more contentious uses of (2)–(4) try to express.

Viewed from this perspective, his pre-Tarskian position is comprehensible. He realized that sentences, in some sense, represent the world and so have meanings and truth conditions. But, he thought, there is no way to state or formulate what this comes to. Wrongly thinking that claims about how language represents the world are ruled out, he looked for something in the conceptual neighborhood that might capture what he took claims about meaning and truth to be confusedly trying to capture. This is the

[30] Carnap (1932 [1959]), p. 62. By "elsewhere" he means *The Logical Syntax of Language*.

route that led him to identify *the analysis of meaning* with *the analysis of the logical syntax of language*. The practice of analysis in this sense required translation of ordinary language into a formal language (Russellian logical form), plus the specification of logical properties and relations of sentences of the formal language in terms of syntactic notions like formal derivability.

The following passages from Carnap (1934b) illuminate this systematic program of linguistic (indeed syntactic) analysis.

> On the basis of the concept "entailment" [formal deducibility] one can define the following classification of propositions [sentences] which is fundamental to the logic of science. A proposition is called *analytic* . . . if it is an entailment of every proposition. . . . A proposition [sentence] is called *contradictory* if any proposition [sentence] at all is its entailment. A proposition [sentence] is called *synthetic* if it is neither analytic nor contradictory. An analytic proposition [sentence] is true in every possible case and therefore does not state which case is at hand. A contradictory proposition [sentence] on the contrary says too much, it is not true in any possible case. A synthetic proposition [sentence] is true only in certain cases, and states therefore that one of these cases is being considered,—all (true or false) statements of fact are synthetic.[31]

> And now we come to the principal concept of the logic of science, the concept of the content of a proposition [sentence]. Can this central concept . . . be formulated purely formally also? . . . [W]hat . . . do we want to know when we ask concerning the content or meaning of a proposition [sentence] S? We wish to know what S conveys to us; what we experience through S, what we take out of S. In other words: we ask what we can deduce from S; more accurately: what propositions [sentences] are entailments of S which are not already entailments of any proposition [sentence] at all, and therefore declare nothing. We define therefore: by the *content* (Gehalt) of a proposition [sentence] S we understand the class of entailments from S which are not analytic. Thereby the concept "Gehalt" is connected to the syntactical concepts defined earlier; it is then also a syntactic, a purely formal concept. . . . Thus the defined concept "Content" corresponds completely to what we mean when we (in a vague manner) are accustomed to speak of the "meaning" (Inhalt) of a proposition [sentence]; at any rate, insofar as by "meaning" or "sense" of a proposition [sentence] something logical [as opposed to psychological—e.g., what one "thinks of" or "imagines"] is meant.[32]

This is the sense of the *analysis of content* Carnap had in mind when he identified the job of philosophy as the analysis of the syntax of the language of science—where by science he meant any systematic attempt to state empirical facts (including those formulated in ordinary language).

[31] Carnap (1934b), pp. 11–12.
[32] Ibid., pp. 12–13.

As previously indicated, the burden of Carnap (1932) was to explain the means by which the impediments to the analysis of the language of science—metaphysics and normative theory—were to be eliminated by being characterized as meaningless. As we have seen, word meaning was to be given by specifying the contents (nonanalytic entailments) of elementary sentences in which words appear. Whenever a word W is definable in terms of another expression E, the contents of elementary sentences containing W will match those of corresponding sentences containing E. According to Carnap, the process of defining words in terms of other words continues until we reach primitive observational vocabulary, the meanings of which are given by the fact that the elementary sentences in which they occur are direct reports of sense experiences.

> In this way every word of the language is reduced to other words and finally to the words which occur in the so-called "observation sentences" or "protocol sentences." It is through this reduction that the word acquires its meaning. For our purposes we may ignore entirely the question concerning the content and form of the primary sentences (protocol sentences). . . . At times the position is taken that [these] sentences . . . speak of the simplest qualities of sense and feeling . . . others incline to the view that basic sentences refer to total experiences and similarities between them [the *Aufbau*]; a still different view has it that even the basic sentences speak of things. *Regardless of this diversity of opinion, it is certain that a sequence of words has a meaning only if its relations of deducibility to the protocol sentences are fixed . . . and similarly, that a word is significant only if the sentences in which it may occur are reducible to protocol sentences*. [33]

For Carnap at this time, every meaningful empirical term was either itself an observation term or an observationally definable term. Meaningless terms found in metaphysical and normative theories don't satisfy this condition. One example not mentioned by Carnap is 'good' as G.E. Moore understood it in *Principia Ethica*. Another, which he does mention, is 'God', when used to refer to a being beyond our experience.[34]

Carnap discusses examples of what he takes to be meaningless but grammatical sentences some of which are made up entirely of meaningful words. These are found only in natural languages, which, to his dismay, allow syntactically well-formed expressions to which coherent meanings can't be assigned on the basis of the meanings of their parts. His examples include (1)–(3).

1. God exists.
2. I think, therefore I am.
3. Caesar is a prime number.

[33] Carnap (1932 [1959]), p. 63, my emphasis.
[34] Ibid., p. 66.

First consider (1) and (2). About them, Carnap says:

> [I]t has been known for a long time that existence is not a property. . . . But it was not until the advent of modern logic that full consistency on this point was reached: the syntactical form in which modern logic introduces the sign for existence is such that it cannot, like a predicate, be applied to signs for objects, but only to predicates. . . . An existential statement does not have the form "*a* exists" (as in "I am," i.e. "I exist"), but "there exists something of such and such a kind." . . . The second error lies in the transition from "I think" to "I exist." If from the statement "P(a)" ("a has the property P") an existential statement is to be deduced, then the latter can assert existence only with respect to the predicate P, not with respect to the subject *a* of the premise. What follows from "I am a European" is not "I exist," but "a European exists." What follows from "I think" is not "I am" [or "I exist"] but "there exists something that thinks."[35]

Here, Carnap recycles old mistakes. Adverting to the Frege-Russell analysis of quantification, which was a genuine advance, he repeats some of the errors that have, unfortunately, often been associated with it. In chapters 2, 8, and 12 of volume 1 I argued that although it was never shown that existence isn't a property of objects (expressed by the predicate 'exists'), there is good reason to think it isn't a property of Fregean concepts or of Russellian propositional functions, as Frege and Russell seem to suggest. One wonders what Carnap (who says both that existence is *not* a property and that the symbol for it is '∃') takes the semantic function of '∃' to be. One also wonders how it can be denied that the claim *that a exists* could fail to follow from the claim *that a thinks*, or *that a is a European*. After all, one can't *think* or *be a European*, if one doesn't exist, even though one can be *dead*, *admired*, or *designated by a name* even if, like Socrates, one doesn't exist but once did. So, it seems that 'Rudolf exists' *does* follow from 'Rudolf thinks' and 'Rudolf is a European', in which case what he expressed by 'I exist' does too.[36]

Next consider (3), which Carnap claims to be meaningless.

> "Prime number" is a predicate of numbers; it can neither be affirmed nor denied of a person. Since [3] . . . does not assert anything and expresses neither a true nor a false proposition, we call this word sequence a "pseudo-statement." The fact that the rules of grammatical syntax are not violated easily seduces one at first glance into the erroneous opinion that one still has to do with a statement, albeit a false one. But "a is a prime number" is false iff a is divisible by a natural number different from a and from 1.[37]

The passage doesn't make it clear why Carnap characterizes the necessary and sufficient conditions for falsity of the claim that a is a prime number

[35] Ibid., p. 74.
[36] See Soames (2014), pp. 62–64, 395–97, 598–604; also Salmon (1987, 1998).
[37] Carnap (1932 [1959]), p. 68.

in the way that he does. Let H be the set of all trios of human beings. Far from being either meaningless or neither true nor false, the sentence '*Since H isn't a number, H isn't a prime number*' seems to be *true*. Indeed, we can say something similar about people. '*Since people aren't numbers, Caesar isn't a number, and hence he isn't a prime number*' also seems true, rather than meaningless.

Later in the article, Carnap indicates why he thinks what he does.

> Another very frequent violation of logical syntax is the so-called "*type confusion*" of concepts. . . . We have here a violation of the rules of the so-called theory of types. An artificial example is the sentence we discussed earlier: "Caesar is a prime number." Names of persons and names of numbers belong to different logical types, and so do accordingly predicates of persons (e.g. "general") and predicates of numbers ("prime number").[38]

Carnap is here referring to Russell's theory of types, understanding it as Russell wished it to be understood—not simply as a convenient method to block set-theoretic paradox in *Principia Mathematica*, but as a constraint on the very intelligibility of talk about numbers in particular and classes in general. Here too, Carnap recycles a mistake.

As I argued in chapter 10 of volume 1, *ontological interpretations* of Russell's higher-order quantification in *Principia Mathematica*—quantification over classes or nonlinguistic propositional functions—allow the derivations to go through with maximum simplicity, without paradox. But the statements about classes or propositional functions ruled out by the simple type theory that goes with this interpretation of the quantifiers are not plausibly taken to be meaningless or unintelligible. On the contrary, although the segmentation of the formulas of the logical language into discrete types avoids paradox, it does so by artificially limiting expressive power. Indeed, it appears impossible to describe the principles governing the entire hierarchy without saying things that the type hierarchy does not allow one to say. For this reason, what is presented as a higher-order system of *logic* can seem, when interpreted in the now standard objectual way, to be a particular theory with its own subject matter, rather than what Russell desired—*a general logical framework governing reasoning about any subject*.

Realizing this, he introduced substitutional elements into his discussion of quantification in *Principia Mathematica*. This, I argued, underlies his infamous *no-class theory* and *ramified* (as opposed to simple) *theory of types*. Because the type theory flowing from the substitutional interpretation incorporates genuine constraints on the coherence and intelligibility of *this kind of quantification*, the type-theoretic constraints can be defended as nonarbitrary and purely logical. However, when Russell's system is interpreted in this way it is not strong enough for his ambitious logicist

[38] Ibid., p. 75.

purposes; it also threatens important elements of his general philosophical logic.[39] In short, the substitutional interpretation isn't a good bargain for Russell. It's also not a good bargain for Carnap. But without it, he no longer has a compelling reason to think that (3) is a meaningless but grammatical sentence.

The important point about (3) is not, of course, whether it is meaningless and neither true nor false, as Carnap contends, as opposed to being obviously false. The important point is that he needs to justify his claim that large domains of discourse about religion, metaphysics, and morality are both truth-valueless and cognitively meaningless. For this he doesn't have to rely on anything as specialized as one version of the theory of types. He already has a general justification—namely, his thesis that every meaningful sentence is either empirically verifiable, empirically falsifiable, knowable solely in virtue of meaning, or refutable solely on that basis. The real challenge is to make this thesis precise enough to be evaluated, and then to make it palatable when so formulated. Although the challenge would, as we will see, prove formidable, in 1932 Carnap seemed to think the following summary was enough.

> (Meaningful) statements are divided into the following kinds. First there are statements which are true solely in virtue of their form. . . . They say nothing about reality. The formulae of logic and mathematics are of this kind. They are not in themselves factual statements, but serve for the transformation of such statements. Secondly, there are negations of such statements ("*contradictions*"). They are self-contradictory, hence false in virtue of their form. With respect to all other statements the decision about truth or falsehood lies in the protocol sentences. They are therefore (true or false) *empirical statements* and belong to the domain of empirical science. Any statement one desires to construct which does not fall within these categories becomes automatically meaningless. Since metaphysics does not want to assert analytic propositions, nor to fall within the domain of empirical science, it is compelled to employ words for which no criteria of application are specified and which are therefore devoid of sense, or else to combine meaningful words in such a way that neither an analytic (or contradictory) statement nor an empirical statement is produced. In either case pseudo-statements are the inevitable product.[40]

3. HANS HAHN: THE LINGUISTIC THEORY OF THE A PRIORI

In 1933 Hans Hahn, one of the founding members of the Vienna Circle, published the pamphlet "Logik, Mathematik und Naturerkennen," later translated and reprinted (in part) as "Logic, Mathematics, and the

[39] See Soames (2014), chapter 10, section 5, and also Soames (2015b), section 3.
[40] Carnap (1932 [1959]), p. 76.

Knowledge of Nature," in Ayer (1959). In it he argues that apriority is entirely the result of linguistic convention, and so provides no knowledge of the world. This is presented as the solution to the problem for empiricists of making room for logical and empirical knowledge. Hahn explains why he took this to be a threat to empiricism.

> But . . . empiricism faces an apparently insuperable difficulty: how is it to account for the real validity of logical and mathematical statements? Observation discloses to me only the transient, it does not reach beyond the observed; there is no bond that would lead from one observed fact to another, that would compel future observations to have the same result as those already made. The laws of logic and mathematics, however, claim *absolute universal* validity. . . . [Because of this] the propositions of logical and mathematics . . . cannot be derived from experience . . . In view of the tremendous importance of logic and mathematics in the system of our knowledge, empiricism, therefore, seems to be irrevocably refuted.[41]

But what, exactly, is the worry? What empiricist claim is supposed to be threatened by a priori knowledge? Is it the claim that there is no a priori knowledge? Is it the claim that there is no a priori knowledge *about the world*? Is it the claim that a priori knowledge is needed to derive new empirical knowledge from antecedently known empirical propositions?

Hahn elaborates what he finds threatening:

> The usual conception, then, may be described roughly as follows: from experience we learn certain facts, which we formulate as "laws of nature"; but since we grasp by means of thought the most general lawful connections (of a logical and mathematical character) *that pervade reality*, we can *control* nature on the basis of facts disclosed by observation to a much larger extent than it has actually been observed. *For we know in addition that anything which can be deduced from observed facts by application of logic and mathematics must be found to exist*. According to this view the experimental physicist provides knowledge of laws of nature by direct observation. The theoretical physicist thereafter *enlarges this knowledge* tremendously by thinking, *in such a way that we are in a position to assert propositions about processes that occur far from us in space and time and about processes which, on account of their magnitude or minuteness, are not directly observable but which are connected with what is directly observed by the most general laws of being, grasped by thought, the laws of logic and mathematics*. . . .
>
> Nevertheless, we are of the opinion that this view is entirely untenable. For on closer analysis it appears that the function of thought is immeasurably more modest than the one ascribed to it by this theory. The idea that thinking is an instrument *for learning more about the world than has been observed, for acquiring knowledge* of something that has absolute validity always and

[41] Hahn (1933 [1959]), pp. 149–50.

> everywhere in the world, an instrument for *grasping general laws of all being*, seems to us wholly mystical.[42]

Here Hahn states his opposition to a view of the roles of observation and a priori derivation in obtaining empirical knowledge. Although there are clearly errors in the view he opposes, his opposition to two propositions deserves closer scrutiny.

P1. A priori truths of logic and mathematics are sometimes used to derive new a posteriori knowledge of the world from a posteriori knowledge we already possess.
P2. The laws of logic and mathematics are a priori truths, which are the most general laws governing everything in nature.

Hahn expresses his opposition to P1, in the continuation of the last cited passage.

> Just how should it come to pass that we could predict the necessary outcome of an observation before having made it? Whence should our thinking derive an executive power, by which it could compel an observation to have this rather than that result? Why should that which compels our thoughts also compel the course of nature?[43]

This is confused. P1 raises the question *Is it possible to know p on the basis of empirical evidence, to know a priori that if p is true, then q is true, by virtue of deducing the empirical truth q from p, and thereby come to know q?* When Hahn asks, rhetorically, "How could we predict the *necessary* outcome of an observation before having made it?" he must be taking his question to be different from the question "How could we predict the outcome of an observation before having made it?" What, then, is 'necessary' doing in his question? No one thinks that in using the necessary truth *if p is true, then q is true* to derive q from p, we thereby confer necessity on q, compelling it to occur, as it were. Stripped of this error, Hahn's question loses its rhetorical force.

Nevertheless, he does have a related, more interesting worry about P1. He seems to think that if q is a proposition that could be false, and we don't already know that q is true, then instead of *coming to know q by deriving it from p*, our derivation merely exposes the fact that *we didn't know p in the first place*. Rather than extending empirical knowledge, deduction threatens it.

> We said that the usual view was roughly this: experience teaches us the validity of certain laws of nature, and since our thinking gives us insight into the most general [logical and mathematical] laws of all being, *we know that likewise*

[42] Ibid., p. 151, my emphasis.
[43] Ibid., p. 151.

> *anything deducible from these laws of nature by means of logical and mathematical reasoning must be found to exist*. We now see this view is untenable; for thinking does not grasp any sort of laws of being. *Never and nowhere, then, can thought supply us with knowledge about facts that goes beyond the observed*. . . . Let us ask ourselves, e.g., what was involved in the computation of the position of the planet Neptune by Leverrier! Newton noticed that the familiar motions, celestial as well as terrestrial, can be well described in a unified way by the assumption that between any two mass points a force of attraction is exerted which is proportional to their masses and inversely proportional to the square of their distance. *He could not pronounce this law as a certainty*, *but only as a hypothesis. For nobody can know that such is really the behavior of every pair of mass points—nobody can observe all mass points*. . . . Leverrier's calculations made people aware that the assertion of the law of gravitation implies that at a definite time and definite place in the heavens a hitherto unknown planet must be visible. People looked and actually saw the new planet—the hypothesis of the law of gravitation was confirmed. *But it was not Leverrier's calculation that proved that this planet existed, but the looking, the observation*. This observation could just as well have had a different result . . . in which case the law of gravitation would not have been confirmed and one would have begun to doubt whether it is really a suitable hypothesis. . . . Indeed, this is what actually happened later.[44]

Here Hahn confuses an uncontentious platitude with a dubious philosophical thesis. Of course, if q is a logical/a priori/necessary consequence of p, and q turns out to be false, then p is also false, and hence not known. But the mere fact that an agent A who might otherwise count as knowing p doesn't already know q—and so doesn't already know that q isn't false—isn't enough show that A doesn't know p either. If knowing p *doesn't* guarantee knowing all its logical/a priori/necessary consequences already, then the fact that A doesn't know q is compatible with A's knowing p, even though q is a consequence of p. Indeed, the position seemingly suggested by Hahn's remarks about Newtonian gravitation—that no hypothesis capable of being known entails the occurrence of observational events that can't all actually be observed—is far too strong, since it rules out much scientific knowledge.

I doubt that Hahn would be moved by this critique, because I suspect his real position is that it is *impossible* for agents *not* to know, believe, and assert all logical/a priori/necessary consequences of what they know, believe, or assert. If so, then P1 is, as he contends, false because anyone who knows, believes, or asserts p already knows, believes, or asserts q, without doing any deduction at all. This view is suggested by his tractarian assumptions (i) that all and only logical truths are necessary and a priori

[44] Ibid., pp. 160–61, my emphasis. Hahn goes on to cite observational evidence leading to the replacement of Newtonian gravitation by Einsteinian gravitation.

truths, (ii) that logical truths are tautologies and so say nothing, and (iii) that the conjunction of a tautology with an empirical truth p says nothing more and nothing less than p.

Establishing (i) is Hahn's aim in writing the article, whereas (ii) and (iii) are suggested by several passages. Here is a representative remark about tautologies.

> To sum up: we must distinguish two kinds of statements: *those which say something about facts* and those which merely *express* the way in which the rules govern which application of words to facts depend upon each other. Let us call statements of the latter kind *tautologies*: they say nothing about objects and are for this very reason certain, universally valid, and irrefutable by observation.[45]

Not only do tautologies *say nothing about objects*, they also don't *say anything about linguistic rules*. If they did, they would be empirical and so, in principle, refutable, which they are not. Rather they "merely express"—Wittgenstein would say "show"—something about rules. So tautologies don't say anything, and conjoining one with an empirical proposition p doesn't change what is said by p.

In the presence of (i)–(iii), all we need to get the conclusion that it is impossible to know, believe, or assert p without knowing, believing, or asserting all logical/a priori/necessary consequences of p is the pair of trivial assumptions (iv) and (v): (iv) propositions—thought of as bearers of truth and falsehood—are the objects of knowledge, belief, and assertion; (v) knowledge, belief, and assertion distribute over conjunction—i.e., if ⌜A knows/believes/ asserts that S&R⌝ is true, then ⌜A knows/believes/asserts that S⌝ and ⌜A knows/believes/asserts that R⌝ are also true. This, I contend, is Hahn's reason for rejecting P1.

For example, in the following passage he suggests that whenever we explicitly assert a proposition, we implicitly assert all logical/a priori/ necessary consequences of it.

> Thus we have arrived at something fundamental. . . . [I]n asserting the two propositions "object A is either red or blue" and "object A is not red," I have implicitly *already asserted* "object A is blue." *This is the essence of so-called logical deduction*. . . . A person who refused to recognize logical deduction would *not* thereby manifest a different belief from mine about the behavior of things, but he would refuse to speak about things according to the same rules I do. . . . What logical deduction accomplishes, then, is this: it makes us aware of all we have implicitly asserted.[46]

The problem is not with Hahn's example, but with his promiscuous generalization of it. We do implicitly assert all the relevant and trivially obvious

[45] Ibid., p. 155. My emphasis.
[46] Ibid., 156–57, my emphasis.

consequences of things we explicitly assert. For example, if I say to you "Mary solved the problem," and someone asks you what I said in a context in which the fact that the problem was solved is more important than the identity of the one who solved it, then your report "Soames said/asserted that someone solved the problem" is true. Similarly in Hahn's example, if I say both "A is either red or blue" and "A is not red," I can truly be reported to have asserted that A is blue. What is *not* acceptable is the conclusion he draws from this observation—namely that agents assert all *logical/a priori/necessary* consequences of what they explicitly assert. Suppose Sam happens to assert each of the premises of Gödel's first incompleteness theorem without drawing any further conclusions. Surely, we *cannot* truly say "Sam asserted that all omega-consistent extensions of the first-order theory Q of arithmetic are incomplete." Nor is Fred truly described by ⌜Fred asserted that S⌝ *for every sentence S*, simply because he mistakenly asserts that first-order Peano arithmetic is complete. Yet these incredible results are suggested by Hahn's unsupported generalization.

Even Hahn is uncomfortable with his thesis.

> [L]ogical propositions . . . being purely tautologous, and logical deductions . . . being nothing but tautological transformations, have significance for us because we are not omniscient. Our language is so constituted that in asserting such and such propositions we implicitly assert such and such other propositions—but *we do not see immediately all that we have implicitly asserted in this manner*. It is only logical deduction that makes us conscious of it.[47]

Here it sounds like we may *not* believe some propositions that are logical/a priori/necessary consequences of propositions we assert and believe. If so, then Hahn is admitting that the attitudes are *not* closed under such consequence. However, he immediately muddies the water, making it unclear whether deduction generates a belief in a consequence of our premises *that we didn't previously believe*, or whether it merely brings a belief to consciousness that we already had.

> I assert, e.g., the propositions "the flower which Mr. Smith wears in his buttonhole, is either a rose or a carnation," "if Mr. Smith wears a carnation in his buttonhole, then it is white," "the flower which Mr. Smith wears in his buttonhole is not white." *Perhaps I am not consciously aware that I have implicitly asserted also "the flower Mr. Smith wears in his buttonhole is a rose"; but logical deduction brings it to my consciousness.*[48]

The equivocation recurs in different forms throughout his discussion.

> The propositions of mathematics are of exactly the same kind as the propositions of logic: they are tautologous, *they say nothing at all about the objects* we

[47] Ibid., p. 157, my emphasis.
[48] Ibid., p. 157, my emphasis.

> wish to talk about, but *concern* only the manner in which we want to speak about them. The reason why we can assert apodictically with universal validity the proposition: 2+3 = 5 . . . is that by "2+3" we mean the same as by "5". . . . We *become aware* of meaning the same . . . by going back to the meanings of "2," "3," "5," "+," and making tautological transformations *until we see* that "2+3" means the same as "5." . . . [E]very mathematical proof is a succession of such tautological transformations. Their utility is due to the fact that, for example, *we do not by any means see immediately that we mean by "24 x 31" the same as by "744"; but if we calculate the product . . . we recognize that . . . what we mean after the transformation* [calculation] is still the same as what we meant before it, until finally we become *consciously aware* of meaning the same by "744" as by "24 x 31."[49]

Here the claim seems to be (a) that we always *mean the same thing* by pairs of sentences that are necessary and a priori consequences of each other, and by pairs of singular terms used to construct identity statements that are necessary, a priori truths, but (b) *we often don't know that we mean the same thing until we perform the necessary deductions*. The discussion here is interesting because statements about what we mean by expressions or sentences are always contingent and empirical. Thus, it might seem that Hahn is suggesting that we learn which mathematical sentences are true by doing the required calculations, even though we already know the single empty triviality that all the necessarily true sentences are used to express.

But this too is untenable. Anyone who has learned the rules governing the symbols of a mathematical language, and is able to use them in conjunction, will, by virtue of knowing the conjunction of the rules, know something that entails a correct statement of the truth conditions, *S is true iff P*, for each sentence S of the language.[50] So, if the attitudes are closed under logical/necessary/a priori consequence, the agent will know that *S is true iff P*. When the sentence replacing 'P' is itself logically, metaphysically, and epistemically necessary, the agent will already know the triviality it expresses, and so also know that S is true. Thus, if the attitudes are closed under logical/necessary/a priori consequence, there will be no metalinguistic knowledge to be gained by doing the proofs. Perhaps that is why Hahn equivocates, saying that we "finally become consciously aware" of meaning the same thing by the necessarily equivalent expressions, rather than saying that we finally *learn*—i.e., come to believe—that they mean the same thing. Such equivocation doesn't help. Either Hahn takes the attitudes to be closed under necessary/a priori/logical consequence or he doesn't. If he doesn't, he has no argument against P1. If he does, he has

[49] Ibid., p 158, my emphasis.
[50] 'S' is here used as a metalinguistic variable over sentences; 'P' is a sentential schematic letter.

adopted a highly counterintuitive thesis without arguing for it. The underlying logic of his discussion suggests the latter.

With this, we turn to his criticism of P2.

P2. The laws of logic and mathematics are a priori truths, which are the most general laws governing everything in nature.

Although Hahn agrees that the laws of logic and mathematics are a priori, he takes them to be true in virtue of linguistic convention alone, and hence not about anything, let alone the things existing in nature. Here is a sample.

> We learn, by training I am tempted to say, to apply the designation "red" to some of these objects, and we stipulate that the designation "not red" be applied to all other objects. On the basis of this stipulation we can assert with absolute certainty the proposition that there is no object to which both the designation "red" and the designation "not red" is applied. It is customary to formulate this briefly by saying that nothing is both red and not red. This is the law of contradiction. And since we have stipulated that the designation "red" is to be applied to some objects and the designation "not red" to *all* other objects, we can likewise pronounce with absolute certainty the proposition: everything is either designated as red or as "not red," which it is customary to formulate briefly by saying everything is either red or it is not. This is the law of the excluded middle.[51]

Later, he adds, "It is this convention about the use of negation which is expressed by the laws of contradiction and of the excluded middle."[52]

What are the conventions about redness and negation to which Hahn alludes? Perhaps they are R and R~.

R1. For all x, if x looks like this . . . , then 'red' applies to x.
R1~. For all x, if x does not look like this . . . , then 'not red' applies to x.

However, this pair of stipulations can't be the linguistic conventions that govern our understanding of both 'red' and 'not'—since in order to understand the stipulation, one must already have mastered negation. There is also a related problem. Let's grant that it follows from R1 and R1~ that everything is such that either 'red' or 'not red' applies to it. How is it supposed to follow that there is nothing to which they both apply? If the application of 'red' and 'not red' is completely determined by R and R~, it won't follow. Perhaps this can be avoided by substituting R2 and R2~ for R1 and R1~.

R2. For all x, if x looks like this . . . , then 'red' applies to x and if 'red' applies to x, then x looks like this . . .

[51] Ibid., p. 153.
[52] Ibid., p. 156.

R2~. For all x, if x does not look like this . . . , then 'not red' applies to x and if 'not red' applies to x, then x does not look like this . . .

But this doesn't help much, since to get the desired results we must assume that the logical constants 'if, then', 'and', 'not', and 'all' are already understood, and that the logical laws governing them are already in place. So understood, the stipulations don't explain the logical laws; they presuppose them. Hahn has nothing else to offer.[53]

4. SCHLICK'S FOUNDATION OF KNOWLEDGE

Schlick (1934 [1959]) discusses *knowledge, certainty, truth, confirmation*, and what logical empiricists called "protocol statements," which were taken to be the terminus of empirical verifications or falsifications. The article, which stirred a storm of controversy, is one of the most interesting and illuminating pieces of the period. Although the concepts discussed were central to logical empiricism, the article, and the reactions to it, revealed the depth of confusion about them that existed among Schlick and his colleagues.

The central question is whether certainty is required as the foundation of empirical knowledge. Schlick thinks it is. Suppose S is empirically meaningful. Then, S must be verifiable or falsifiable. If S is not only verifiable but verified, then it is a candidate for being known to be true. But what is it to verify a statement? Though Schlick offers no definition, he maintains that verification rests on knowledge of observational facts, expressed by "protocol statements," which he characterizes as follows:

> statements which express the *facts* with absolute simplicity, without any moulding, alteration, or addition, in whose elaboration every science consists, and which precede all knowing, every judgment regarding the world. It makes no sense to speak of uncertain facts. . . . If we succeed therefore in expressing the raw facts in "protocol statements," without any contamination, these appear to be the absolutely indubitable starting points of all knowledge.[54]

For Schlick, protocol statements are the terminus of all empirical verification. Without verification by protocols, we don't know any empirical statements to be true.

What about the protocols themselves? What is it for them to be *certain*, and why must they be? We may assume one thing; in order to provide *evidence*, protocols must themselves be *known*. What, then, is the relationship

[53] The notion *truth by convention* and the linguistic doctrine of the a priori is discussed in greater detail in chapter 10.
[54] Schlick (1934 [1959]), 209–10.

between knowledge and certainty? Perhaps knowledge requires certainty. After all, when we take ourselves to know something, we are reluctant, if challenged, to say "I'm not sure it's true" or "It might not be." So, if by saying "p is certain" one is saying "I'm sure p is true," or "p must, given my evidence, be true," in the sense in which one who knew p would standardly refuse to accept their negations, then there is a way of understanding *certainty* such that to know p is to be certain of p. To be certain of a protocol in this sense is simply to *know* it, and thus to be *sure*, given the perceptual experience that is one's evidence for it, that it is true.

But there are also other ways of understanding certainty. Suppose I know p. Suppose also that q and r are statements that are logically independent of p, but which, if they were true, would undermine my knowledge of p—either by falsifying p (if q were true) or by undermining the evidence on which my knowledge of p is based (if r were true). In such a case, my knowledge of p is, in a certain way, *dependent* on ~q and ~r; if either were shown not to be true, I wouldn't know p. Since my knowledge of p is compatible with the truth of the claim *If ~q or ~r weren't true, I wouldn't know p*, one might think, my knowledge of p is dependent on the "assumption" or "hypothesis" that ~q and ~r are true. Some philosophers might (wrongly) add that if my knowledge of p depends in this way on such assumptions or hypotheses, then I must also *know* those assumptions or hypotheses—in which case p won't itself qualify as a protocol statement. Rather, these philosophers would insist, when p is a genuine protocol, one's knowledge of p is independent of all other assumptions or hypotheses, and hence is *absolutely certain*.

Schlick was such a philosopher. Here is one of his examples.

> There appears in a book a sentence which says . . . that N.N. used such and such an instrument to make such and such an observation. One may under certain circumstances have the greatest confidence in this sentence. [Schlick would say that one may know it.] Nevertheless, it and the observation it records, can never be considered *absolutely* certain. For the possibilities of error are innumerable. . . . Indeed the assumption that the symbols of a book retain their form even for an instant and do not "of themselves" change into new sentences is an empirical hypothesis, which as such can never be strictly verified. . . . This means . . . that protocol statements, so conceived, have in principle exactly the same character as all the other statements of science: they are hypotheses, nothing but hypotheses. They are anything but incontrovertible.[55]

Here we are told that something I may be presumed to know—*namely that NN (or even SS) observed that so-and-so* a moment ago—is *uncertain* because there are "assumptions" which, if they were true, would falsify that

[55] Ibid., p. 212. In speaking of protocol sentences, *so conceived,* Schlick has in mind the conception of protocols as then advocated by Carnap and Neurath.

proposition (or undermine my claim to know it). We are also told that any empirical statement that is uncertain in this sense is merely a *hypothesis* because it is not "incontrovertible." For Schlick, if a statement is a hypothesis, then it is *not* a genuine protocol statement but rather a statement that itself requires verification by protocols.

Why does he think that genuine protocol statements must be *absolutely certain*? Because he implicitly reasons as follows: (i) All empirical statements must be verified by experience if they are to be known. (ii) Call any empirical statement an *empirical hypothesis* iff one's knowledge of it is dependent (in the sense just discussed) on the truth of other (logically independent) empirical statements. (iii) No one can know any empirical hypothesis p without verifying, and hence coming to know, the empirical statements on which one's knowledge of p depends. (iv) So, if any empirical statements are knowable, not all empirical statements can be empirical hypotheses—for if they were, verification would either fail to terminate, or become circular (which it can never be). (v) Since many empirical statements are knowable, there must be some empirical statements knowledge of which is not dependent on the truth of any logically independent statements. (vi) These protocol statements are shown to be true by directly comparing them with the facts of experience, without any need to consider the truth or falsity of logically independent statements. (vii) All empirical verification consists in the judgments expressed by these statements.

Schlick's protocol statements, which don't include statements about the observable properties of physical objects, are statements about "immediate perceptual experience." But they aren't just any such statements. Statements that function as protocol statements for me are never about the perceptual experiences of others; nor are they about my own perceptual experience at any time after the moment of perceptual judgment in which I accept them.[56] According to Schlick, they can't be written down at all.

> [T]he function of the statements about the immediately experienced itself lies in the immediate present . . . they have no duration . . . the moment they are gone one has at one's disposal in their place inscriptions, or memory traces, that can play only the role of hypotheses.[57]

Although Schlick's conception of protocol statements sounds mysterious, it is possible to make some sense of it.

He calls his true protocol statements "confirmations." Apart from being synthetic, he thinks they are closely analogous to analytic truths.

> I cannot raise the question whether I can ascertain the correctness of an analytic statement. For to understand its meaning and to note its *a priori* validity

[56] Ibid., pp. 218–19.
[57] Ibid., pp. 222.

> are in an analytic statement *one and the same* process. In contrast, a synthetic assertion is characterized by the fact that I do not in the least know whether it is true or false if I have only ascertained its meaning. . . . The process of grasping the meaning is here quite distinct from the process of verification.
>
> There is but one exception to this. . . . "Confirmations" . . . are always of the form "Here now so and so," for example . . . "Here yellow borders on blue," . . . "Here now pain". . . . What is common to all these assertions is that *demonstrative* terms occur in them which have the sense of a present gesture. . . . What is referred to by such words as "here," "now" . . . cannot be communicated by means of general definitions in words, but only by means of them together with pointings or gestures. "This here" has meaning only in connection with a gesture. In order therefore to understand the meaning of such an observation statement one must simultaneously execute the gesture, one must somehow point to reality.
>
> In other words: I can understand the meaning of a "confirmation" only by, and when, comparing it with the facts, thus carrying out that process which is necessary for the verification of all synthetic statements. . . . However different therefore "confirmations" are from analytic statements, they have in common that the occasion of understanding them is at the same time that of verifying them. . . . In the case of a confirmation it makes as little sense to ask whether I might be deceived regarding its truth as in the case of a tautology. Both are absolutely valid. However, while the analytic, tautological, statement is empty of content, the observation statement supplies us with . . . genuine knowledge of reality.[58]

The first thing to note is that Schlick's "confirmations" are similar to Russell's sense-data statements circa 1910.[59] A Russellian *logically proper name* was a singular term the meaning of which was its referent. His only examples of such were indexicals and demonstratives. For x to be the referent of A's use of such a term, Russell held that A must be *acquainted* with x, which meant, in effect, that A couldn't be mistaken about x's existence or nature. Because sense data were perceptual appearances, to know their nature was to know their perceptible properties—which were all and only those they appear to have. So, if p is a statement about the perceptible properties of certain of A's sense data, then (i) A can be absolutely certain of p when entertaining it, and (ii) p can be entertained only by A, and only when A is perceiving those particular sense data. Schlick's "confirmations" were, in effect, Russell's sense data statements by another name.

That said, the following section of the passage cited above is troubling.

> What is referred to by such words as "here," "now" . . . cannot be communicated by means of general definitions in words, but only by means of them

[58] Ibid., pp. 224–25.
[59] See Soames (2014), chapter 8 for discussion.

> together with pointings or gestures. "This here" has meaning only in connection with a gesture. In order therefore to understand the meaning of such an observation statement one must simultaneously execute the gesture, one must somehow point to reality.

Although it makes sense to speak of *communicating* with others by *gesturing at* or *pointing to* publicly perceivable things, it doesn't make sense to speak of *communicating* with others by *gesturing at* or *pointing to* one's own private sense data. Thus, Schlick faces a dilemma.

On the one hand, he could drop talk of gestures, pointing, and communication, and insist that the facts reported by "confirmations" are always confined to one's own private experiences. In doing this, he might preserve the certainty of protocol statements, but only at the cost of losing their capacity of verifying the statements of science. Since the construction of scientific theories is a collective effort, their verification must also be. No set of perceptual experiences of a single agent—let alone experiences of an agent at a single moment—is sufficient to verify a significant theory. Nor do individual agent-time verifications sum in a way that provides knowledge. If separate verifications of different agents (at different moments) were all there was to verification, then no one would verify any significant scientific statement or have any scientific knowledge. Hence, this way of remedying the infelicity of the cited passage is unacceptable.

On the other hand, Schlick could take "confirmations," even those expressed using demonstratives, to be about publicly perceivable things at which one can point or gesture. The best way of doing would be to swap talk of *protocol sentences* for talk of *protocol propositions*, identifying the latter with *uses of sentences* containing indexicals and demonstratives to predicate properties and relations of the referents of one's uses of those expressions. It would then follow that an agent's use, at time t and place p, of the indexical sentence 'The object here is hot now' to predicate the property of *being hot at t* of a certain designated stone S was a protocol proposition, pp1, that could be entertained only at t and p. Of course, in entertaining and accepting pp1, an agent would also entertain and accept a representationally identical but cognitively distinct proposition pp2 that predicates *being hot* of S at t, no matter how S and t are identified or cognized—which can be entertained at any time.[60] If an agent A located at p and t entertains the two propositions because A perceives S to be hot, then A's rational confidence in them will be the same at t. However, in many cases it will be natural for A's rational confidence in pp2 to drop over time (e.g., in cases in which all A has to go in is A's memory of t). If he wished, Schlick could express this by saying that only pp1 was a genuine "confirmation." But it could *not* generally be maintained that confirmations, when understood in this way, are *indubitable* or *absolutely certain*. Most properties and relations

[60] See Soames (2015b) for details.

predicated of intersubjectively observable objects and events involved in the verification of scientific theories are *not* the sort that indubitably apply to their predication targets. For this reason, even the verifying "confirmations" would lose their privileged status.

Hence, Schlick's view that for all empirical propositions p, one can know p only if one has evidence for p that is expressed by propositions that are not only known to be true, but also *absolutely certain* (in his sense), can't be correct. Among other things, this means rejecting step (iii) of my earlier reconstruction of his implicit reasoning—*No one can know any empirical hypothesis p without verifying, and hence coming to know, the empirical statements on which one's knowledge of p depends.* There is no absurdity in maintaining that A knows an empirical proposition p by virtue of knowing the verifying evidence for it provided by q, even though there are propositions on whose truth A's knowledge depends that A doesn't know to be true. Schlick didn't see this.

In addition to misunderstanding the relationship between knowledge, certainty, and confirmation, Schlick also appears to have misunderstood the relationship between confirmation and truth. His initial discussion linking the two comes in response to the Carnap-Neurath view that no empirical statements are "certain," that all "hypotheses" are capable of being verified or falsified, and that it is a matter of theoretical convenience which empirical statements are taken to be "protocols." Here is how Schlick connects truth and confirmation.

> For us it is self-evident that the problem of the basis of knowledge [confirmation by protocol statements] is nothing more than the question of the criterion of truth. Surely the reason for bringing in the term "protocol statement" in the first place was that it should serve to mark out certain statements by the truth of which the truth of all other statements comes to be measured. But according to the view [of Carnap and Neurath] . . . this measuring rod would have shown itself . . . relative. . . . But what remains at all as a criterion of truth? Since the proposal is not that all scientific statements must accord with certain definite protocol statements, but rather that all statements shall accord with one another, with the result that every single one is considered as, in principle, corrigible, truth can consist only in a *mutual agreement of statements*. . . . This view . . . is well known from the history of philosophy. In England it is usually called the "coherence theory of truth," and contrasted with the older "correspondence theory of truth." . . . [A]ccording to the traditional one [the correspondence theory] the truth of a statement consists in its agreement with the facts, while according to the other, the coherence theory, it consists in its agreement with the system of other statements.[61]

[61] Schlick (1934 [1959]), pp. 213–14. The views of Carnap and Neurath to which Schlick refers are those in Carnap (1932/1933b) and Neurath (1932/1933).

What started as a dispute about knowledge, certainty, and confirmation here morphs into a dispute about the nature of truth. Schlick, who takes himself to be defending the correspondence theory, makes the standard objection to the coherence theory—*taken as a theory of what truth is*. He argues (i) that the *agreement* with other statements required by the coherence theory in order for a statement S to be true can only be the *consistency* of S with the other statements; but (ii) that consistency isn't sufficient for truth because there are many different individually consistent systems that are inconsistent with each other.[62] Returning to his concern with a *criterion of truth*, he argues that the only alternative to the failed coherence theory is to recognize some statements—of immediate observation—the truth of which must be held fixed against all contingencies. These are to be used to define what it is for the others to be true.

> Thus, the coherence theory is shown to be logically impossible; it fails altogether to give an unambiguous criterion of truth, for by means of it I can arrive at any number of consistent systems of statements which are incompatible with one another. The only way to avoid this absurdity is not to allow any statements whatever to be abandoned or altered [in the face of recalcitrant experience], but rather to specify those that are to be maintained [come what may], to which the remainder have to be accommodated.[63]

How, one wonders, did what started out as a dispute about whether empirical confirmation and knowledge requires a basis in certainty, as Schlick conceived it, get transformed into a dispute about what truth is? To say of an empirical statement S that it is true is *not* to say of it that it has been confirmed or that it eventually will be. One might think that for S to be true is for it to be possible to confirm S. But what is here meant by 'possible'? Not either *metaphysical* or *epistemic possibility*, since we don't want to call a statement 'true' that is actually false, but is also either a statement that is true and confirmable at some possible world-state, or a statement the falsity of which is consistent with everything we know. Nor could we take 'possible' to mean *logical possibility*, since when S is false, the claim that it is confirmed usually isn't contradictory. One might be tempted define *true statements* as those that would be confirmed, if we were able to gather enough evidence, but only if by *enough evidence* one meant something other than *all the evidence needed to show them to be true*. Since there are no obvious, non-question-begging candidates for filling this role, it is unlikely that Schlick implicitly relied on a formulation of this sort. The fact that he had no analysis of counterfactual conditionals, or the concepts needed to give one, makes it all the more unlikely that

[62] Although this is correct, Schlick might have made a more fundamental point— since consistency is itself defined in terms of truth, it can't be used in defining what truth is.
[63] Schlick (1934 [1959]), p. 216.

such reasoning was responsible for his turning a dispute about confirmation into a dispute about truth.

The most likely source of the trouble was, I think, a certain reading of the *Tractatus*, which Schlick and other members of the Vienna Circle had accepted by 1930, and onto which Schlick, among others, grafted a verificationist element. According to this reading, an elementary proposition is true iff it pictures an atomic fact; the truth of every other proposition is *defined* by its agreement, or disagreement, with atomic propositions. This was not a doctrine about confirmation; it was a doctrine about what the truth of these different types of propositions consists in. Since tractarian elementary propositions are logically independent of one another, Schlick concluded that judging such a proposition to be true didn't require any assumptions about the truth or falsity of any other propositions. For that reason, he thought, elementary propositions must be capable of being known with absolute certainty. To deny this, as he took Carnap and Neurath to do, was, in his mind, to turn the tractarian theory of what truth is—correspondence with reality for elementary propositions and coherence with elementaries for non-elementary propositions—into a complete, and disastrously unmoored, coherence theory of truth. Since this was unacceptable, Schlick needed a conception of elementary propositions that explained how we can be absolutely certain of the truth of a certain kind of synthetic statement. The result was his conception of "confirmations"—i.e., uses of sentences that are statements about immediate perceptual experience the grasping of which is sufficient for their (conclusive) verification. For these, he thought, there is no gap between truth and confirmation.

What about empirical statements that aren't "confirmations"? For Schlick, their verification or falsification is always provided by "confirmations." Is their truth, or falsity, *defined* by their agreement or disagreement with (actual or possible) "confirmations"—as it should be, if he is still adhering to the tractarian model? He doesn't tell us. One might suspect that, if he were to accept that their truth, or falsity, *is* so defined, he would be a radical phenomenalist. But he wouldn't have agreed. Rather, he would have dubiously claimed, as he does in Schlick (1932/33), that phenomenalism is a meaningless metaphysical thesis that attempts to solve what is in fact only a pseudo-problem—the classical problem of the reality of the external world.

5. HEMPEL: TRUTH, CONFIRMATION, AND CERTAINTY

In January of 1935, Carl Hempel published "On the Logical Positivists' Theory of Truth"—partly to respond to Schlick, partly to defend Carnap and Neurath, and partly to chronicle recent changes in Carnap's and Neurath's thoughts about truth and confirmation. After giving the usual descriptions of correspondence and coherence theories of truth, Hempel

characterizes the logical positivists as having gradually moved from a correspondence theory based on the *Tractatus* to a "restrained coherence theory."[64] He notes that for Wittgenstein, the truth of "atomic statements" consists in their correspondence with facts, that non-atomic statements are truth functions of atomics, and hence that the truth or falsity of non-atomic statements consists in their relations to atomic statements. Hempel then reports that Neurath believed that no statement can be "compared" with facts, apparently because he, like Wittgenstein, believed that we can't meaningfully describe the relationship between language and the world. Taking Neurath to mean that "each [scientific] statement may be combined or compared with each other statement . . . [b]ut statements are never compared with 'reality' or 'facts'," Hempel attributes a coherence theory of truth to Neurath.[65]

Carnap is described as sharing the view that talk of the relationship between statements and facts is metaphysical nonsense. Because of this, Hempel says, Carnap sought to avoid all such talk, while leaving the rest of the tractarian conception of language in place. His solution was to single out certain statements as never needing proof because they express "the result of a pure immediate experience without any theoretical addition."[66] Hempel describes the substitution of such "protocol statements" for Wittgenstein's atomic statements as "the first step in abandoning Wittgenstein's theory of truth."[67] Hempel takes it to be axiomatic that in substituting protocols for Wittgenstein's atomic statements, Carnap was adopting a form of the coherence theory of truth. Since the tractarian theory of truth was, except for the correspondence account of atomic truth, a coherence theory, Hempel takes Carnap's amputation of the correspondence account of atomic statements to have left him a coherence theory. But this characterization is superficial. Without an answer to the question "In what does the truth of protocol sentences consist?" Carnap had no theory of truth. Because his early position didn't preclude the answer "By representing the objects experienced as having properties they actually do have," Carnap hadn't yet decisively rejected correspondence theories of truth.

According to Hempel, the next step in Carnap's evolution away from the *Tractatus* involved giving up the idea that all meaningful statements are truth functions of atomic statements. The crucial examples are universal generalizations.

> A general statement is tested by examining its singular consequences [Hempel is here thinking of universal generalizations.] But as each general statement

[64] Hempel (1935), p. 49.
[65] Ibid., pp. 50–51.
[66] Ibid., p. 51.
[67] Ibid., p. 51.

> determines an infinite class of singular consequences, it cannot be finally and entirely verified: a general statement is not a truth function of singular statements, but it has in relation to them the character of an *hypothesis*. The same fact may be expressed as follows: a general law cannot be formally deduced from a finite set of singular statements.[68]

This is puzzling. It's true that no empirical universal generalization is a logical consequence of any finite set of its instances. It is also true that no such generalization is formally provable from, or conclusively verified by, any such set. But finitude isn't the issue. Analogous results hold in standard versions of the predicate calculus no matter what cardinality the set of instances of a universal generalization has. Moreover, in the *Tractatus* universal generalizations *can be* truth functions of infinite sets of statements.[69] So, what Hempel here tells us seems merely to be that Carnap and Neurath didn't adopt the tractarian model of quantification.

Because the nature of truth is at issue, we need to know what it is for a universal generalization to be true. Hempel doesn't say.

> Each finite set of statements admits an infinite series of hypotheses [different generalizations] each of which implies all the singular statements referred to [their instances]. So, in establishing the system of science, there is a conventional moment; we have to choose between a large quantity of hypotheses which are logically equally possible, and in general we choose one that is distinguished by formal simplicity.[70]

Since no set of instances of a universal generalization logically entails it, one may ask, "When, if ever, does knowledge of them confirm the generalization, and so *justify* us in taking it to be true?" Hempel seems to answer that it is a matter of convention. Suppose that's right. Still, it doesn't tell us what the *truth* of a universal generalization consists in. If, as one would suppose, it consists in *all the instances* being true, then it will be possible for us to be justified in believing a universal generalization that is false.[71] Surely this trivial observation doesn't threaten the tractarian conception of truth.

The last step identified by Hempel in the purported evolution away from the *Tractatus* is the elimination of any class of statements "which are conceived to be ultimate [final steps in verification] and not to admit of any doubt."[72] For Carnap and Neurath, even protocol statements may sometimes require further empirical verification. Having made this point, Hempel says, "Obviously, these general ideas imply a coherence theory of truth."[73] Why

[68] Ibid., p. 52.
[69] See chapter 3.
[70] Hempel (1935), p. 52.
[71] Here we have to allow for instances to include uses of formulas in which a free occurrence of a variable can be used to designate any object whatsoever. See chapter 3 for discussion.
[72] Hempel (1935), p. 53.
[73] Ibid., p. 54.

is that obvious? It may be obvious, given the views of Carnap and Neurath, that for every empirical proposition p, there are situations in which knowing p depends on the truth of logically independent propositions that themselves require confirmation. But this is a thesis about knowledge, not truth. Forget the *Tractatus* for a moment. Consider instead a version of the correspondence theory that takes all truth to be representational accuracy. According to it, for a proposition p to be true is (i) for p to represent something as being a certain way (e.g., as bearing a certain property or standing in a certain relation) and (ii) for the thing to be that way (to bear that property or stand in that relation). This theory of truth is compatible with Carnap's and Neurath's views about knowledge and verification.

The problems I have found with Hempel's discussion of the implications of Carnap's and Neurath's views of *knowledge, certainty, verification,* and *protocol statements* for views about *truth* stem largely from the fixation of all three philosophers on the *Tractatus*, which inhibited their exploration of other ways of cashing out the correspondence idea. The tractarian collapse of necessity and apriority into uninformative triviality (which all three accepted) also played a role, precluding the recognition of informative necessary truths about the representational properties of propositions. But the foremost problem inherited from the *Tractatus* was the doctrine, discussed in chapter 4, that the relationship between language and the world in virtue of which uses of language are meaningful cannot meaningfully be stated in language. If one didn't believe that, but instead believed that the relationship can be described, then one could contemplate correspondence theories that take truth to be accuracy in representation. Unfortunately, our three philosophers were not in this position at the end of 1934.

This is evident in the response Hempel gives to the question "If the views of Carnap and Neurath in 1934 entail a coherence theory of truth, which such theory is entailed?" Which of the many different, but equally coherent, systems of statements must an empirical statement agree with in order to be true? Hempel recognizes the problem and struggles to solve it.

> As Carnap has shown, each non-metaphysical consideration of philosophy belongs to the domain of Logic of Science, unless it concerns an empirical question and is proper to empirical science. And it is possible to formulate each statement of Logic of Science as an assertion concerning certain properties and relations of scientific propositions only. So also the concept of truth may be characterized in this formal mode of speech, namely . . . as *a sufficient agreement between the system of acknowledged protocol-statements and the logical consequences which may be deduced from the statement and other statements which are already adopted*.[74]

A few pages later, he elaborates the italicized characterization of truth a bit.

[74] Ibid., p. 54, my emphasis.

> The system of protocol statements which we call true, and to which we refer in every day life and science, may only be characterized by the historical fact that it is the system which is actually adopted by mankind, and especially by the scientists of our cultural circle; and the "true" statements in general may be characterized as those which are sufficiently supported by the system of actually adopted protocol statements.[75]

The characterization of truth here gestured at is clearly *not* truth as we understand it in science or everyday life. The main difference between the two is given by the contrast between (i) and (ii):

(i) The claim *that p is true*, as ordinarily understood, is epistemically and metaphysically equivalent to p itself.
(ii) The claim *that p agrees with observational and other statements that have been accepted by scientists (and others) in our cultural circle* is neither necessarily, epistemically, or even materially equivalent to p.

Because of this difference, the notion $truth_H$ that Hempel defines can't play the roles we require of a notion of truth. For example, in the presence of classical logic—which the logical empiricists accepted—$truth_H$ fails to vindicate both the move from p to the claim that p is true, and the move from the claim that p is true to p. For suppose that the claim that p is true can be derived from p. Then, by classical logic, the negation of p can be derived from the claim that p isn't true. This yields a contradiction when $truth_H$ is substituted for *truth* and p is an empirical proposition which—like the proposition *that there is a duplicate of the earth somewhere in the Milky Way*—is such that neither p nor p's negation has been confirmed by scientists in our cultural circle. Hence, $truth_H$ doesn't validate the inference from p to the claim that p is true; instead, p is compatible with the claim that p isn't $true_H$. Nor is the move from the claim that p is true to p validated. Since there is nothing inconsistent about instances of the schema "*Although ~S, the scientists of our cultural circle have accepted that S*," $truth_H$ doesn't allow us to derive p from the claim that p is true.[76] These results are devastating because, when p is a proposition that doesn't itself employ the notion *truth* (or any related notion), the practical and theoretical uses of that notion depend on inferences that $truth_H$ fails to support.[77]

It is interesting that the notion Hempel does define involves a relationship between uses of sentences and certain facts in the world—facts concerning the acceptance of those uses by certain people. But if talk about

[75] Ibid., p. 54.
[76] Instances are obtained by reading '~' as *it is not the case that* and substituting a sentence for the two occurrences of the symbol 'S' in the schema— e.g., *Although it is not the case that all swans are white, scientists of our cultural circle have accepted that all swans are white.*
[77] For discussion of the conceptual uses of our concept of *truth,* see Soames (1999), pp. 22–23, 98–107, 229–31, and also the reprinting of Soames (2003c), "Understanding Deflationism," in Soames (2009b), pp. 337–38.

this relationship between language and the world is meaningful, as it surely is, then talk about other relationships between language and the world should also be. Consider a factual claim about a group of speakers that (i) it is a convention among them to use 'Scott Soames' as a name referring to me and 'is male' as a predicative expression standing for the property *being male* and (ii) that it is also a convention to use sentences ⌜N is P⌝ to predicate the property the predicative expression stands for of the referent of N. It will then follow from these facts plus the nature of predication that uses of 'Scott Soames is male' represent me as being male, and hence are true iff I am male. Indeed, the truth of such a use may *consist in* my being as the use of the sentence represents me to be. To establish this, one need not raise controversial questions about how, or whether, it can be known that I am male, or if it can be known, whether it can be known with certainty. Those are different issues. Unfortunately, all of this was foreign to Carnap, Neurath, and Hempel when the latter wrote his article on truth.

6. REICHENBACH: THE ELIMINATION OF TRUTH

It should, I hope, be clear by now that the tendency to confuse questions of truth with those of knowledge, certainty, and confirmation was widespread among logical empiricists. I will next use the example of another logical empiricist, Hans Reichenbach, to identify one of the factors, beyond the influence of the *Tractatus*, that was responsible for this.

Reichenbach, who studied civil engineering, physics, mathematics, and philosophy, was a student of David Hilbert, Max Planck, Albert Einstein, and Ernst Cassirer, among others. Arguably the leading philosopher of science of his era, he published three major books on relativity theory and related issues in the 1920s. In 1929, he was one of ten logicians, mathematicians, and philosophers mentioned as "sympathetic to the Vienna Circle," in the manifesto written by Carnap, Hahn, and Neurath. In 1930, he was, with Carnap, founding co-editor of *Erkenntnis*. He taught at the University of Berlin from 1926 to 1933, where he became the central figure in the "Berlin Group," which included Kurt Grelling, Kurt Lewin, Richard van Mises, and Carl Hempel (who was Reichenbach's student). In 1933, after the rise of Hitler, he was fired from his university post, after which he moved to Istanbul, where, in 1935, he published *The Theory of Probability*. In 1938 he accepted a teaching position at UCLA, moved to the United States, and published *Experience and Prediction*, which championed the role of probability in science.

In chapter 3 of that book, he summarizes his findings and says the following about truth.

> Throughout the first chapter we entertained the presupposition that propositions about concrete physical facts, which we called observation propositions, are absolutely verifiable [and so absolutely certain]. A more precise analysis

> showed that this conception is untenable, that even for such statements only a weight [i.e., probability] can be determined. With the object of obtaining more reliable statements, we then introduced [sense] impression propositions; throughout the second chapter we upheld the supposition that at least these propositions are capable of absolute verification. We have discovered now that even this is not tenable, that impression propositions also can only be judged by the category of weight. Thus there are left no propositions at all which can be absolutely verified [i.e., of which we can be absolutely certain]. *The predicate of truth-value of a proposition, therefore, is a mere fictive quality; its place is in an ideal world of science only, whereas actual science cannot make use of it*. Actual science instead employs throughout the predicate of weight [probability]. We have shown, in the first place, that this predicate takes the place of the truth-value in all cases in which the latter cannot be determined; so we introduce it for . . . indirect propositions, which remain unverified for all times [i.e., that are never fully and conclusively verified]. We see now that all propositions are, strictly speaking, of the latter type; that all propositions are indirect propositions and never exactly verifiable. *So the predicate of weight has entirely superseded the predicate of truth-value and remains our only measure for judging propositions*. If we, nevertheless, speak of the truth-value of a proposition, this is only a schematization. We regard a high weight as equivalent to truth, and a low weight as equivalent to falsehood.[78]

Reichenbach's thesis is that empirical propositions can never be established with complete certainty—where certainty is not an overwhelming feeling of confidence, but a state in which one's basis for accepting a proposition rationally guarantees its truth. At most, Reichenbach thinks, experience can render a proposition highly probable; nothing can guarantee that further experience won't require one to reject it. Having reached this conclusion, he immediately draws the further, skeptical, conclusion that truth is a fiction which has no place in science. But why does he draw this conclusion? Why should the claim that certainty is unattainable lead one to think that truth is too?

Like other logical empiricists of his day, Reichenbach didn't explain why he linked truth with certainty in this way. But he said enough about how he took truth to be related to probability and confirmation to provide a clue. The following is a reconstruction of a seductive line of thought that I believe influenced Reichenbach and many others.[79]

S1. If the proposition *that P* is empirical, then what one is committed to by virtue of assertively uttering
   a. It is true *that P* / The proposition *that P* is true

[78] Reichenbach (1938), pp. 187–88, my emphasis.
[79] I first explicated this line of thought in chapter 2 of Soames (1999). Capital 'P' is used here and throughout this section as a schematic sentence letter.

is *stronger* than what one is committed to in virtue of assertively uttering b. It is highly probable / confirmed / supported *that P*.

In each case—(a) and (b)—one is expected to have evidence strongly supporting the proposition *that P*. But for (a), this is insufficient, since one is also committed to the proposition *that P*. If that proposition turns out to be untrue, then one who has assertively uttered (a) will have made an error; this is not always so for one who has assertively uttered (b).

S2. Thus, the statement made by uttering (a) is stronger than the one made by uttering (b).

S3. The strongest statement one is warranted in making about any empirical proposition is that it is highly probable, confirmed, or supported. No empirical statement can ever be established with complete certainty; rather, every empirical statement is a more or less probable hypothesis the acceptance of which is a function of its role in our total scientific worldview.

S4. Thus one is never warranted in making the statement expressed using (a). Since empirical truth is unattainable, truth has no legitimate place in empirical science.

Though superficially seductive, this argument is far too sweeping. We know a priori that *P iff it is true that P*. So, if we were never warranted in asserting that *the proposition that P is true*, we would never be warranted in asserting the proposition *that P* (for any empirical proposition). In other words, if scientific methodology excludes truth, then it excludes all empirical propositions. This is a *reductio ad absurdum* of the view.

The argument confuses truth with certainty, taken as the limiting case of high probability. For any proposition p, p is probable to degree n iff the proposition *that p is true* is probable to degree n.[80] Not so with certainty. The probability that *a particular coin will come up heads the next time it's flipped* is, we may assume, .5. Thus, the probability that *it is true that the coin will come up heads the next time it's flipped* is also .5. But the probability that *it is certain that the coin will come up heads the next time it's flipped* is not .5. So truth must be distinguished from certainty.

With this in mind, let us reexamine the steps in the argument. S1 is correct; but S2 isn't. *The act of asserting* that it is true that P by assertively uttering (a) commits one to the proposition asserted—and *also* (perhaps) to having evidence that renders that proposition highly probable, which is, of course, evidence that renders the proposition *that P* highly probable, too. But *what one asserts*—the statement one makes in assertively uttering (a)—doesn't entail anything about one's evidence. It neither entails nor is entailed by the proposition that one has evidence rendering it highly probable that P. Thus S2 is false; neither the statement made by assertively

[80] Here and throughout this section, lowercase 'p' is an objectual variable over propositions.

uttering (a), nor the statement made by assertively uttering (b), is stronger than the other—in the sense of committing one to what the other commits one to, and more.

Step S3 is also problematic. Even if we accept the claim that no empirical proposition can be known with certainty, we must ask what it means for a statement q to be *stronger than* the statement that p is highly probable. If it means that q entails that statement, but not conversely, then S3 is unacceptable. On such a definition, the conjunction of any empirical proposition p with the proposition that p is highly probable will be stronger than its second conjunct. But surely, one is sometimes warranted in asserting both p and the proposition that p is highly probable (even though the former isn't absolutely certain). So, on this interpretation of *strength,* the argument fails.

Perhaps, however, all that is meant by S3 is that no empirical proposition can ever be established with complete certainty. If so, then, it could be maintained that no one is ever warranted in claiming that an empirical proposition p is certain; the most one can claim—about how probable p is—is that p is highly probable. On this interpretation of *strength*, a statement q about an empirical proposition p is stronger than a statement r about p iff q attributes higher probability to p than r does. So understood, S3 need not be contested. Since its notion of strength differs from the one in S1 and S2, the argument equivocates and S4 isn't established.

We have, then, no good argument to support Reichenbach's claim that predications of truth are illegitimate, or that truth is epistemically unattainable. The key to recognizing this is to observe that the claim *that P* is necessarily equivalent to the claim *that it is true that P* (while also being knowable a priori to be so) which, in turn, is necessarily equivalent to the claim *that the proposition that P is true* (while being knowable a priori to be so). Once this is noted, it is obvious that truth is distinct from certainty, and that the intelligibility and legitimacy of truth wouldn't be threatened, even if it were shown that certainty was unattainable. As we will see in chapter 9, Carnap, for one, didn't realize this until after he learned of Tarski's "definition of truth."

CHAPTER 8

Advances in Logic: Gödel, Tarski, Church, and Turing

1.BACKGROUND

1.1. Overview and Chronology

In 1929, when his name appeared on the list of members in "The Scientific Conception of the World" announcing the existence of the Vienna Circle, Kurt Gödel was a PhD student at the University of Vienna. In the next

year, his dissertation was accepted and published as Gödel (1930), the English title of which is "The Completeness of the Axioms of the Functional Calculus of Logic." In it he proved that the first-order predicate calculus is *complete* in the sense that *every logical truth* in a first-order language is *provable* from logical axioms and rules of inference—where a *proof* is a finite sequence of lines, each of which is an axiom or a formula obtainable from earlier lines by the inference rules. It is crucial to this notion of *proof* that one can always decide, merely by inspecting the formula on a line, whether or not it is an axiom, and, if it isn't, whether or not it bears the required structural relation to earlier lines for it to be obtainable from them by the rules. To impose these requirements is to insist that proof be *effectively decidable*. The point of imposing them is to ensure that whether or not something is a proof can always definitively be resolved—thereby forestalling the need to prove that something is a proof.

Gödel's completeness proof shows that all *logical truths* are provable, and every *logical consequence* B of a sentence (or set of sentences) A is provable from A. To understand what this means, one must distinguish *logical* from *nonlogical* symbols. The logical vocabulary of a first-order language consists of one or both of the quantifiers '∀' and '∃' plus some truth-functional connectives, e.g., '&', 'v', '~', '→', and '↔'. (The identity predicate '=' can be treated either as logical or as nonlogical.) A sentence in the language is *logically true* iff it is true, and would be so (i) no matter what (nonempty) domain of objects its quantifiers were taken to range over, and (ii) no matter how its nonlogical vocabulary were interpreted to apply (or not apply) to those objects—i.e., no matter which objects its names designated, which its predicates were true of, or which functions from things in the domain to things in the domain its function signs designated.

Although this conception of logical truths wasn't itself made the subject of precise meta-mathematical investigation until the notion of a *model*, or *interpretation*, of a formal language was formalized in Tarski (1935, 1936), the informal idea was available to Gödel in 1930. Using it, we say that S is *logically true* iff S is true in all models of the language, and that Q is a logical consequence of a sentence (or set of sentences) P iff every model that makes P (or its members) true also makes Q true. Since only the interpretation of the logical vocabulary remains fixed across models, this fits the idea that logic alone, independent of any assumptions about special subject matter, is sufficient to determine logical truth (and consequence).

There are three natural demands one might place on a logical proof procedure. The minimal demand is that every provable sentence be true in all models, and so be a logical truth. This demand is within the reach of any system of logic worthy of the name. The intermediate demand is that the procedure be complete, in the sense that Gödel (1930) proved there to be complete first-order proof procedures. The satisfaction of these two demands ensures the existence of formalizations of first-order logic in which the logically provable sentences coincide with the logical truths. A

third demand one might hope to be satisfied is that there be a *decision procedure* for logical truth—a procedure which, given any sentence S, always decides correctly (in finitely many steps) whether S is, or isn't, logically true. Although the truth-table method is such a procedure for the propositional calculus, it was proved by Alonzo Church in 1936, and independently by his student Alan Turing in 1937, that no similar procedure for the first-order predicate calculus is possible.[1] This result, which will be explained later in the chapter, was a corollary of Gödel's first incompleteness theorem.

That theorem, presented in Gödel (1931), states that any ω-consistent, axiomatizable, first-order theory T of the arithmetic of the natural numbers will be *incomplete*, in the sense that there will be pairs of sentences S and ⌜~S⌝ neither of which is provable in T. As we will see, this result was later strengthened by substituting *consistency* for Gödel's slightly weaker original notion of *ω-consistency*. (A consistent theory is one that never proves both S and its negation.) So, with the full strengthening of Gödel's original result, we have it that every consistent, axiomatizable. first-order theory T that doesn't prove contradictions will fail to prove some truths. Gödel's second incompleteness theorem is presented in the extension Gödel (1932) of Gödel (1931). It states that if an axiomatizable, first-order theory T of arithmetic is consistent, then one of the *truths* that T will be incapable of proving is a certain arithmetical statement C_T that we can prove—*in our metatheoretic reasoning about T*—to be a theorem of T only if T proves no contradictions. Since C_T can be taken "to say" that the arithmetical theory T is consistent, this result is standardly said to show that *no consistent, axiomatizable, first-order theory of arithmetic is capable of proving its own consistency*.

Before conveying more precisely what these results mean, and how they were established, I need to say a few words about formal languages and theories in general, about the particular case of arithmetic, and about the ingenious Gödelian technique of introducing a special convention, when investigating an arithmetical theory T, that allows us to read certain sentences in the language L_T of T as "making claims" about T and the sentences of L_T. In what follows I will use a small font to present matters of detail which, though important to master the technical material, can be skipped or skimmed for those who wish to concentrate of the leading ideas.

1.2. Formal Languages and Theories: Arithmetic

The vocabulary and formation rules of first-order languages distinguish between logical and nonlogical vocabulary.

[1] Church (1936a, 1936b), Turing (1936/37).

Vocabulary

Nonlogical

There will be finitely many names, predicates, and function signs.

Logical

The logical vocabulary includes truth-functional connectives—e.g., '~', '&', '∨', '→', '↔'—plus either or both of '∀', '∃' along with infinitely many variables 'x', 'x_n'. For any variable ν, ⌜∀ν⌝ and ⌜∃ν⌝ are quantifiers. The identity predicate '=' can be treated either as logical or nonlogical. I will here treat it as logical.

Terms

Names and variables are terms; If $t_1 \ldots t_n$ are terms and f is an n-place function sign, then the result of combining them is a term.[2] Nothing else is a term.[3]

Formulas

a. A combination of an n-place predicate with n terms is a formula of the language.[4]
b. If A and B are formulas, so are ⌜(~A)⌝, ⌜(A & B)⌝, ⌜(A ∨ B)⌝, ⌜(A→B)⌝, ⌜(A↔B)⌝.
c. If A is a formula and ν is a variable, then so are ⌜(∀νA)⌝ and ⌜(∃νA)⌝.
d. Nothing else is a formula.

Sentences

a. A sentence is a formula containing no free occurrences of any variable.
b. An occurrence of a variable ν in a formula A is free iff it is not within the scope of any occurrence of a quantifier using ν.
c. The scope of an occurrence of a quantifier is the quantifier itself plus the smallest (complete) formula immediately following it.

An interpretation of a first-order language consists of the choice of a domain of objects, an assignment of members of the domain as referents of the names, an assignment of functions from n-tuples of objects to other objects in the domain to n-place function signs, and an assignment of n-place properties (defined over objects in the domain) to n-place predicates. Although this informal notion of an interpretation was clear enough for Gödel's purposes, it was not fully formalized until Tarski (1936), from which the contemporary notions of a *model* of a language, *denotation in a model*, and *truth in a model* are abstractable.

Model

A *model* M for a first-order language L consists of a nonempty set D plus an assignment of denotations from D to nonlogical symbols of L. M assigns each name n a member o of D, each n-ary predicate P a set of n-tuples

[2] I here abstract away from the precise syntactic manner of combining them, which, includes ⌜$f(t_1 \ldots t_n)$⌝ as one option.

[3] This is a simplifying convenience. Fregean definite descriptions could also be allowed as terms.

[4] Again, I abstract away from the precise syntax of combination.

of elements of D, and each n-ary function symbol f an n-place function f from n-tuples of members of D into D.

Denotation in a Model

The denotation of a variable ν relative to an assignment A of objects in D to the variables of L is the object that A assigns to ν. The denotation of a term ⌜f (t_1 . . . t_n)⌝ relative to an assignment A (of objects to variables) is the object f assigns to the n-tuple of denotations of t_1 . . . t_n relative to A. The denotation of a name n of L (relative to any assignment of objects to variables) is the object M assigns to n. The denotation of an n-place predicate P is the set of n-tuples of D that M assigns to P. Since '=' is here treated as logical, it always denotes the set of all pairs <o,o> of objects of the domain.

Truth in a Model

An atomic formula ⌜Pt_1 . . . t_n⌝ is true in M relative to an assignment A iff the n-tuple of denotations of t_1 . . . t_n in M relative to A is a member of the denotation of P in M.

⌜~Q⌝ is true in M relative to an assignment A iff Q is false (not true) in M relative to A.

⌜Q & R⌝ is true in M relative to A iff Q and R are both true in M relative to A.

⌜Q ∨ R⌝ is true in M relative to A iff either Q or R (or both) are true in M relative to A.

⌜Q→R⌝ is true in M relative to A iff either Q is false or R is true in M relative to A.

⌜Q↔R⌝ is true in M relative to A iff both are true, or both are false, in M relative to A.

⌜∃νQ⌝ is true in M relative to A iff there is an object o in D such that Q(ν) is true in M, relative to an assignment A* that differs at most from A in assigning o to ν.[5]

⌜∀νQ⌝ is true in M relative to A iff for every object o in D, Q(ν) is true in M relative to an assignment A* that differs at most from A except for assigning o to ν.

A sentence S is true in a model M iff S is true in M, relative to all assignments of objects of the domain of M to variables.[6]

These notions can be illustrated using the first-order language LA of arithmetic. Its nonlogical vocabulary consists of the predicate '=', the name '0', the one-place function sign 'S', and the two-place function signs '+' and '× '. The domain of the usual interpretation of LA, often called *the intended model,* is the set of natural numbers; '=' stands for identity, '0' names zero, 'S' stands for the function that assigns each natural number its successor, and '+' and '× ' designate the addition and multiplication

[5] Q(ν) arises from the quantified formula by erasing the quantifier.

[6] This is explained in chapter 9.

functions, respectively. Since every member of the domain is denoted by *a numeral,* i.e., a member of the series '0', 'S(0)', 'S(S(0))', . . . , we can specify *truth* and *denotation* for LA, *in the intended model N*, without relativizing them to assignments of values to variables.[7]

The Intended Model N of the Language of Arithmetic

The domain is the set of natural numbers. '0' denotes zero; '=' denotes the set of pairs the first member of which is the same number as the second; 'S' denotes the successor function; '+' denotes the addition function; '×' denotes the multiplication function.

Denotation of Terms

If t_1 and t_2 are terms, the denotations of $\ulcorner S(t_1)\urcorner$, $\ulcorner(\, t_1 + t_2)\urcorner$, and $\ulcorner(\, t_1 \times t_2)\urcorner$ are, respectively, the successor of the denotation of t_1, the sum of the denotations of t_1 and t_2, and the product of those denotations.

Truth in Arithmetic

$\ulcorner t_1 = t_2\urcorner$ is true iff t_1 and t_2 denote the same natural number.

Truth conditions for truth negations, conjunctions, etc. are characterized in the usual way.

$\ulcorner \exists v Q\urcorner$ is true iff there is a number n designated by a numeral *n*, such that the sentence $Q(n)$ is true.

$\ulcorner \forall v Q\urcorner$ is true iff for every number n there is a numeral $\underline{n}$ designating n, such that $Q(\underline{n})$ is true.

Next we consider formalized first-order theories. A formal theory consists of a decidable set of axioms, which are sentences of the language of the theory, plus, in some cases, a finite number of definitions. Theorems are sentences provable from the axioms and definitions. Given the completeness proof for first-order logic, these are all and only logical consequences of the axioms and definitions. One such theory is Peano arithmetic, PA. Its (nonlogical) axioms are A1, A2, A3, and all instances of the schema A4, where an instance is the result of replacing the occurrences of 'F(x)' in A4 with any formula $\mathcal{F}$ of LA in which 'x' is the only variable that occurs free, and replacing 'F(0)' with the result of substituting the numeral '0' for all free occurrences of 'x' in $\mathcal{F}$.[8]

A1. $\sim\exists x\,(0 = S(x))$

Zero isn't a successor of anything (any natural number).

A2. $\forall x\,(\sim(x = 0) \rightarrow \exists y\,(y = S(x)))$

Everything (every natural number) except zero is the successor of something (some natural number).

[7] The official definition of *truth in a model* remains so relativized. In the special case of the intended model N, the unrelativized characterization is equivalent to it.

[8] The familiar axiom that every natural number has a successor is built into the notation that treats 'S' as a function sign (standing for a totally defined function) rather than as a two-place predicate.

A3. $\forall x \forall y \forall z\, [(S(x) = z \,\&\, S(y) = z) \rightarrow x = y]$
If the successor of x is the successor of y, then x is identical with y (i.e., no two numbers have the same successor).

A4. $[(F(0) \,\&\, \forall x\, ((F(x) \rightarrow F(S(x))) \rightarrow \forall x\, F(x)]$
If F is true of zero and whenever F is true of a number it is also true of its successor, then F is true of every number.

The definitions of PA are:

D+. $\forall x\, \forall y\, [(x + 0) = x \,\&\, (x + S(y)) = S(x + y)]$
For any natural numbers x and y, the sum of x and 0 is x, and the sum of x and the successor of y is the successor of the sum of x and y.

D*. $\forall x\, \forall y\, [(x \times 0) = 0 \,\&\, ((x \times S(y)) = (x \times y) + x)]$
For any natural numbers x and y, the result of multiplying x times zero is zero, and the result of multiplying x times the successor of y is the sum of x and the result of multiplying x times y.

Proofs in PA are finite sequences of lines, each of which is either (i) an axiom or definition, or (ii) the result of applying a rule of inference to finitely many earlier lines in the sequence. Since there is an effective procedure for deciding whether a given line in the sequence meets (i) or (ii), the class of proofs in PA is effectively decidable. Since every theorem appears on the last line of some proof, it is possible, in principle, to recursively enumerate the theorems of PA—i.e., to construct a list into which every theorem, and only theorems, will eventually appear. Although the list is infinite, we can describe a way constructing it guaranteeing that *if S is a theorem*, we can always determine that it is in a finite series of steps. Since there is no upper bound on how many steps may be required to do so, it is *not* guaranteed that the sentences that aren't theorems of PA can always be determined by consulting the list. (This last result, which was proved by Church and Turing, is explained in section 5.)

We can recursively enumerate the formulas of LA, called (one-place) *predicates*, in which exactly one variable has free occurrences. Given some fixed way of alphabetizing them, we can assign each predicate a unique number k representing its place on the infinite list. Given a definition of truth in the intended model N, we may ask, for each F_k, which, if any, natural numbers F_k is *true of*. In this way, we implicitly associate each predicate with its *extension* (at N)—i.e., the (possibly empty) set of the natural numbers of which it is true. Next we define the set X of natural numbers n such that *F_n is not true of n*. This is the set of sentences such that $\ulcorner F_n(\underline{n}) \urcorner$ is not true in LA (where $\underline{n}$ is the numeral designating n, and $\ulcorner F_n(\underline{n}) \urcorner$ results from substituting $\underline{n}$ for all occurrences of the variable with free occurrences in F_n). This set *can't*, on pain of contradiction, be the extension of any predicate, since if it were, its index k* would, by definition, be a member of X iff it isn't a member of X. This can't be, so X is not the extension of any predicate of LA.

This isn't surprising. We know from Cantor's theorem that there are *uncountably* many sets of natural numbers.[9] Since there are only *countably* many predicates of LA, there are only *countably* many sets of natural numbers that are extensions of them. Thus, there are *uncountably many* sets of natural numbers that are not extensions of any predicates of LA. Nevertheless, the identification of X as among them leads to a pair of interesting questions. Just as we can recursively enumerate the (one-place) predicates of LA, so we can recursively enumerate the sentences of LA, i.e., the formulas with no free occurrences of variables. The questions are:

(i) Is there a way of enumerating the sentences of LA so that the set of indices of true sentences is the extension of a predicate of LA?
(ii) Is the set of true sentences of LA recursively enumerable?

It is a consequence of Gödel's incompleteness theorems that the answer to both questions is "No." The first step toward seeing this is to grasp his system of assigning numbers to formulas and to sequences of formulas that count as proofs in formal theories of arithmetic. This makes it possible to use predicates of LA to encode properties of its own sentences, including their provability or unprovability in arithmetical theories, thus allowing one to take some sentences to encode claims about their own provability or unprovability.

1.3. *Gödel Numbering*

This system of numerical encoding must satisfy three conditions: (i) every formula, and every finite sequence of formulas, is assigned a single Gödel number; (ii) no number is the Gödel number of more than one formula or sequence; (iii) there is an effective procedure for deciding, given any formula or sequence, what its Gödel number is, and also for deciding, given any natural number (a) whether it is the Gödel number of any formula or sequence, and (b) if it is, what it is the Gödel number of. The following is a system meets these conditions.

First each individual symbol in LA is assigned a unique number, as follows:

(	2
)	3
x_1	4
x_2	40
x_3	400
.	
.	
.	
0	5
S	6

[9] Soames (2014), chapter 7, section 2.

+	7
×	8
=	9
~	10
∨	11
&	12
→	13
↔	14
∃	15
∀	16

To assign a number to a formula F, we assign a prime number to each successive place in F occupied by an occurrence of a symbol. The first place is assigned the number 2, the second is assigned 3, the third is assigned 5. Each succeeding place is assigned the next higher prime. Each prime is then raised to the power n, where n is assigned to the symbol occupying that place. To arrive at the Gödel number of F, one multiplies all these primes raised to their respective powers. So, the Gödel number of '$(x_1 = S(x_2))$' is $2^2 \times 3^4 \times 5^9 \times 7^6 \times 11^2 \times 13^{40} \times 17^3 \times 19^3$. Although such calculations are (to say the least) laborious, they can always, in principle, be done. It is a theorem of algebra that for every number n, there is a unique sequence of powers of primes such that n = the product of that sequence. Consequently, every formula of LA will be assigned a unique, and effectively identifiable, Gödel number.

To find out whether an arbitrary natural number n is the Gödel number of a formula, and if so, which, we divide by 2 as many times as possible, then by 3 as many times as possible, then by 5, and so on. We continue until our calculation comes out without remainder. We then have the sequence of powers of primes that yields n. All that remains is to consult the table mapping symbols of LA onto their numerical codes to determine (i) whether the power to which each prime in the sequence is raised is the number assigned one of the symbols of LA, and, (ii) if it is, whether the corresponding sequence of symbols is a formula of LA. Since this is always decidable, the system of Gödel numbering satisfies our requirements.

Finally, we can apply a version of the technique to finite sequences of formulas. To do so, we just treat each formula in a sequence of k formulas as occupying a position, the first of which we associate with the kth prime raised to the Gödel number of the formula occupying it, and so on, assigning successive positions successive primes raised to the Gödel numbers of the formulas occupying them. The Gödel number of the sequence of formulas is the product of these. In this way, each sequence is assigned exactly one Gödel number (distinct from the Gödel number of any single formula), no two sequences are assigned the same number, and the mapping from sequences to numbers and from Gödel numbers back to sequences is decidable. Since all proofs in PA are finite sequences of formula, each is thereby assigned a Gödel number.

In the years after Gödel's initial system of numbering was developed, simpler ones have been offered. Here is one.

(	1
)	19
~	2
∨	29
&	299
→	2999
↔	29999
∃	3
∀	39
x_1	4
x_2	49
x_3	499
.	
.	
.	
0	5
S	6
+	7
×	79
=	8

In this system, the Gödel number of a formula is the number denoted by the Arabic numeral resulting from concatenating (i.e., writing down one after another) the Arabic numerals of the Gödel numbers of the individual symbols that make it up. For example, the Gödel number of $\ulcorner\sim\exists x_1\ (0 = S(x_1))\urcorner$ is 2,341,586,141,919. The Gödel number of a sequence of lines is the number denoted by the Arabic numeral resulting from concatenating the Arabic numerals of the Gödel numbers of each of the formulas in the sequence taken in order. So long as the system of assigning formulas and sequences of formulas Gödel numbers is 1:1—each expression getting its own unique numerical code—and decidable, it doesn't matter what system one uses.

1.4. Definability, Provability, and Truth

Call a set S of natural numbers *definable in arithmetic* iff S is the extension (at the intended model N) of a (one-place) predicate of LA—i.e., iff there is a predicate that is *true of* all and only the members of S (at N). Similarly, a set of n-tuples of natural numbers is *definable in arithmetic* iff that set is the extension of an (n-place) predicate of LA. Using this notion together with a system of Gödel numbering, we can interpret certain sentences of LA as making claims about formulas and sequences of formulas the Gödel numbers of which are the official subject matter of those sentences (interpreted at N). Let P be a (one-place) predicate of LA—which, by definition, is a formula that contains free occurrences of exactly one variable v. Call the (unique) sentence that results from substituting the numeral

that designates the Gödel number of P for all free occurrences of ν in P the *self-ascription* of P. Now consider the relation R that holds between a pair of numbers n and m iff m is the Gödel number of a predicate P of LA and n is the Gödel number of the self-ascription of P. This relation is effectively decidable, since given any pair of natural numbers, we can always determine which expressions of LA, if any, they are Gödel numbers of, and, given any pair of expressions, we can always decide if one is a predicate of LA and the other is its self-ascription. Next, we make use of an important fact about LA: *Every effectively decidable set of, or relation on, natural numbers is definable in it; for every k-place relation R, there is a formula in LA (in which exactly k variables have free occurrences) which is true (at N) of an arbitrary k-tuple of natural numbers iff that k-tuple is an instance of R*. Thus, the relation *y is the Gödel number of the self-ascription of a predicate with Gödel number x* is definable in LA. Let *Self-Ascription x,y* abbreviate a formula of LA that defines it.[10]

Similar reasoning establishes that the relation *x is the Gödel number of a proof in PA of the sentence the Gödel number of which is y* is definable in LA. Let *Proof x,y* be our abbreviation of a formula of LA (in which exactly two variables have free occurrences) that defines it.[11] Next consider the one-place predicate ⌜∃x *Proof x,y*⌝ of LA that results from using the existential quantifier to bind occurrences of the variable corresponding to 'x' in *Proof x,y*. This predicate defines the set of *Gödel numbers of sentences of LA that are provable in the system PA*. Hence, its negation, ⌜~∃x *Proof x,y*⌝, defines the set of *Gödel numbers of sentences of LA that are not provable in PA*. We abbreviate this pair of one-place predicates as *Prov y* and *~Prov y*.

Next consider the one-place predicate (1) that is true of a natural number n iff n is the Gödel number of a one-place predicate of LA the self-ascription of which is not provable in PA.

1. ∃y (*Self-Ascription x,y & ~Prov y)*

The extension of (1) (at N) is the set of Gödel numbers of one-place predicates of LA *that are not provable in PA of their own Gödel numbers*. Finally, we let k be the Gödel number of (1) and $\underline{k}$ be the numeral denoting k. This gives us (2), which is true (at N) iff the self-ascription of (1) is not provable in PA.

2. ∃y (*Self-Ascription* $\underline{k}$*, y & ~Prov y*)

[10] Although, strictly speaking, many formulas of LA define the relation. I take *Self-Ascription x,y* as abbreviating a maximally simple one. If F is such, its conjunction F+ with any truth of LA also defines *self-ascription*. Since its extra complexity is irrelevant, I exclude it (and other irrelevant formulas) from being candidates for *Self-Ascription x,y.*

[11] The point made in the previous footnote applies equally to *Proof x,y.* This point will apply throughout when talking about formulas *defining* sets or relations, or (in our later discussion) *representing them in a theory.*

But (2) is the self-ascription of (1)! Since (2) is true at N iff it is not provable in PA, we can informally take it as predicating *not being provable in PA* of itself. With this, we can generate the simplest version of Gödel's first incompleteness result.

2. SIMPLE GÖDEL INCOMPLETENESS AND GÖDEL-TARSKI INDEFINABILITY

Consider again the formula of LA that (1) abbreviates. It is a predicate of LA that is (provable metatheoretically to be) true of a natural number n iff n is the Gödel number of a sentence that is not provable in the formal theory PA (the theorems of which are encoded by *Prov y*). Next consider the sentence of LA that (2) abbreviates. Since it is the self-ascription of the sentence abbreviated by (1), it is (provable metatheoretically to be) true (at N) iff it is not provable in PA. Assuming that every sentence of LA is true or false (at N), it follows that either the sentence abbreviated by (2) is true and unprovable, or false and provable. *If our formal system PA of arithmetic proves only truths*, it follows that the sentence abbreviated by (2) is true (at N) and unprovable, while its negation is false and also unprovable. So, *if PA proves only truths, then it fails to prove all the truths*; it is *incomplete*, in the sense that there are sentences S and ⌜~S⌝ of LA neither of which is a theorem of PA, despite the fact that one of them is true.

Although we can, given PA, identify which member of this pair is true (at N), this *doesn't* allow us to construct a first-order formalization of arithmetic that is complete. Suppose we add the true sentence of LA corresponding to (2) to the axioms of PA, thereby generating a new theory PA_2 that proves all the truths of PA and more. When we formalize the notion of proof in PA_2, there will be other predicates $Proof_2\ x,y$ and $Prov_2\ y$ of LA—distinct from our original predicates *Proof x,y* and *Prov y*. These allow us to recapitulate our original reasoning about PA to show that PA_2 is also incomplete. In this way, one could, in principle, generate a sequence of first-order theories of arithmetic, each extending its predecessors, all of which are incomplete, despite proving only truths. Every such first-order extension of PA that proves only sentences that are true in the intended model N fails to prove infinitely many such truths. This is the weakest form of Gödel's first incompleteness theorem.

The Gödel-Tarski theorem of the arithmetical indefinability of arithmetical truth is a corollary. It says that there is no predicate LA that is true (at N) of all and only the Gödel numbers of true sentences of LA. For if there were such a predicate *True x*, then there would be a predicate *~True x* that was true of all and only the natural numbers that are *not* Gödel numbers of truths of LA. There would also be a predicate (3a) that was true of all and only those natural numbers m that are Gödel numbers of predicates that are *not true* of their own Gödel numbers.

3a. ∃y (Self-Ascription z,y & ~True y)

Let k be the Gödel number of (3a). Then (3b)—which is the self-ascription of (3a)—is true (at N) iff it is not true (at N).

3b. ∃y (Self-Ascription k, y & ~True y)

Since to say that there is a sentence (3b) of LA that *is true iff it is not true* is a contradiction, the supposition—that there is a predicate (3a)—is false. Since there would be such a predicate if there were a predicate *~True* x of LA that was true of all and only the Gödel numbers of true sentences of LA, no predicate of LA is a truth predicate for LA. Although the set of *provable* sentences of, e.g., PA, is definable in arithmetic, the set of *true* sentences of arithmetic is not.

3. GÖDEL'S FIRST INCOMPLETENESS THEOREM

3.1. Recursive Functions are Representable in the Simple Arithmetical Theory Q

Although the result achieved expresses the intuitive idea behind the first incompleteness theorem, the theorem itself was broader, extending to a class of first-order theories of arithmetic that included, but was not limited to, extensions of PA. Exactly how extensive was not immediately clear in Gödel (1931), but would soon become so. No attempt will be made to trace all historical details of this development. But the eventual reach of the theorem will be specified.[12]

To do so, we let Q be the weak theory the arithmetical axioms of which are Q1–Q7.[13]

Q1 ∀x ∀y (S(x) = S(y) → (x = y))
For all x and y (natural numbers), if the successor of x = the successor of y, then x = y; i.e., no two things (natural numbers) have the same successor.

Q2 ∀x ~(0 = S(x))
Zero isn't the successor of anything (any natural number).

Q3 ∀x (~(x = 0) → ∃y (y = S(x)))
For any (natural number) x, if x isn't zero, then x is the successor of something (some natural number).

Q4 ∀x (x + 0 = x)
For any (natural number) x, the sum of x and zero is x.

[12] My statement and proof of the original theorem, though conforming to Gödel's basic idea, is a simplification and generalization of the original that clarifies its scope on the basis of later work noted below.

[13] Q, which goes back to Robinson (1950), is like PA but lacking the induction axioms. Its use in conjunction with Gödel's incompleteness theorem is covered in Tarski, Mostowski, and Robinson (1953).

Q5 $\forall x \forall y\ (x + S(y) = S(x + y))$
For any (natural numbers) x and y the sum of x and the successor of y is the successor of the sum of x and y.

Q6 $\forall x\ (0 \times x = 0)$
For any (natural number) x, the product of x and zero is zero.

Q7 $\forall x \forall y\ (x \times S(y) = (x \times y) + x)$
For any (natural numbers) x and y, the product of x and the successor of y is the sum of x and the product of x and y.

The eventual reach of the first incompleteness theorem included all consistent extensions of Q.

GÖDEL'S FIRST INCOMPLETENESS THEOREM (STRENGTHENED VERSION)
There are no consistent, complete, axiomatizable first-order extensions of Q.

I have already said what it is for a theory to be consistent and complete. For theory B to be an extension of theory A is for the theorems of theory B to include all the theorems of theory A (and possibly more). I will say that a theory T is axiomatizable iff the set of its axioms is decidable—i.e., iff there is a purely mechanical decision procedure which, given a formula F of language of T, will always correctly decide after finitely many steps whether or not F is an axiom of T.

In order to establish Gödel's first incompleteness theorem, we need to recreate, or simulate, *within all consistent, axiomatizable, first-order extensions of Q*, the reasoning used to establish the simple incompleteness result of the previous section. There I used the semantic notion *definability*, plus the fact that *all decidable sets of, or relations on, natural numbers are definable in LA*. This allowed us to construct a sentence of LA that is (provable metatheoretically to be) true (at N) iff it is not provable in PA, which we could think of as "saying" that it is not provable (in PA). On the assumption that PA proves only truths, it followed that neither it nor its negation is provable in PA, and thus that PA is incomplete. Since certain consistent extensions of Q prove some sentences that are *false* in N (the domain of which consists of all and only the natural numbers), I can't make the same appeal to *truth* in the general version of the first incompleteness theorem that I did in establishing the simple incompleteness result. Since *definability* makes crucial use of *truth*, we need a nonsemantic, proof-theoretic counterpart of definability that can play a role in our generalized reasoning like that played by *definability* in our earlier reasoning.

As we have seen, a set S of k-tuples of natural numbers is *definable* in LA iff there is a k-place predicate of LA that is *true* (at N) of all and only the members of S. The proof-theoretic counterpart of *definability* is *representability* in Q (and any consistent extension). A set S of k-tuples of natural numbers is *representable* in Q iff there is a k-place predicate $P(v_1 \ldots v_k)$ of LA (i) which is *provable* (in Q) of a k-tuple $n_1 \ldots n_k$ of natural numbers iff that k-tuple is a

member of S and (ii) the negation, $\sim P(v_1 \ldots v_k)$, of which is *provable* (in Q) of any other k-tuple of natural numbers.[14] A k-place predicate P is *provable of* $n_1 \ldots n_k$ iff substituting the numeral $\underline{n}_i$ denoting n_i for each occurrence of v_i in P_k is a theorem of Q (similarly for the negation of P).[15]

Just as every decidable set of, or relation on, natural numbers—including the set of pairs such that *m is (or is not) the Gödel number of a proof in Q of the sentence the Gödel number of which is n*—is definable by a predicate of LA, so every such set or relation—including the set of pairs such that *m is (or is not) the Gödel number of a proof in Q of the sentence the Gödel number of which is n*—is *representable* in Q.[16] In other words, we have (4a)–(c).

4a. If m is the Gödel number of a proof in Q of the sentence the Gödel number of which is n, then this is provable in Q—i.e., $Q \vdash$ Proof $\underline{m},\underline{n}$. If m is not the Gödel number of a proof in Q of the sentence the Gödel number of which is n, then, we have $Q \vdash \sim$Proof $\underline{m},\underline{n}$.

b. If n is the Gödel number of a sentence of LA that is provable in Q, then this is provable in Q—i.e., $Q \vdash \exists x$ Proof x, $\underline{n}$ (otherwise put, $Q \vdash$ Prov $\underline{n}$).

c. If n is the Gödel number of a self-ascription of a (one-place) predicate with Gödel number m, then this is provable in Q—i.e., $Q \vdash$ Self-Ascription $\underline{n}$, $\underline{m}$. If n is not the Gödel number of a self-ascription of a (one-place) predicate with Gödel number m, then, that is provable in Q—i.e., $Q \vdash \sim$*Self-Ascription* $\underline{n}$, $\underline{m}$.

It is useful, before using these facts, to say a bit more about representability in Q. The official definition of representability of an n-place total function (i.e., one defined on every argument) in Q is as follows:

> An n-place function f is representable in Q iff there is a formula $A(x_1 \ldots x_n, x_{n+1})$ of LA such that for any natural numbers $m_1 \ldots m_n, m_{n+1}$, if $f(m_1 \ldots m_n) = m_{n+1}$, then $Q \vdash \forall x_{n+1} (A(\underline{m}_1 \ldots \underline{m}_n, x_{n+1}) \leftrightarrow x_{n+1} = \underline{m}_{n+1}$. In such a case, $A(x_1 \ldots x_n, x_{n+1})$ represents f in Q.

This definition is equivalent to:

> An n-place function f is representable in Q iff there is a formula $A(x_1 \ldots x_n, x_{n+1})$ of LA such that for any natural numbers $m_1 \ldots m_n, m_{n+1}$, if $f(m_1 \ldots m_n) = m_{n+1}$, then $Q \vdash A(\underline{m}_1 \ldots \underline{m}_n, \underline{m}_{n+1})$ and $Q \vdash \forall x_{n+1} (A(\underline{m}_1 \ldots \underline{m}_n, x_{n+1}) \rightarrow x_{n+1} = \underline{m}_{n+1}$.

[14] The need for two conditions here—where definability required only one—arises from the fact that whereas for every sentence S of LA, either S or its negation, is true, it is not the case that for an arbitrary formal theory T, either S or its negation is a theorem of T. Since we want the theorems of T to settle the membership of any set representable in T, we need both (i) and (ii) above.

[15] Think of the variables as ordered by their numerical subscripts, so that the free variable with the least subscript is replaced by the numeral of the first element of the k-tuple of natural numbers, and so on.

[16] This is shown in chapter 14 of Boolos and Jeffrey (1974).

Since for all natural numbers n and m, if $n \neq m$, then $Q \vdash \sim(\underline{n} = \underline{m})$, it will follow that if f is representable in a consistent extension Q+ of Q, then (i) $A(\underline{m_1}, \ldots \underline{m_n}, \underline{k})$ will be a theorem of Q+ iff $k = m_{n+1}$ iff $f(m_1, \ldots m_n) = m_{n+1}$ and (ii) $\sim A(\underline{m_1}, \ldots \underline{m_n}, \underline{k})$ will be a theorem of Q+ iff $k \neq m_{n+1}$ iff $f(m_1, \ldots m_n) \neq m_{n+1}$. So, if a computable function is representable in a consistent extension Q+, then we can read off its values for any arguments from the theorems of Q+.

The crucial fact about consistent extensions of Q is the all computable functions are representable in them. We will say that a set S of n-tuples is representable iff its characteristic function (which assigns 1 to members of S and zero to nonmembers) is representable. As noted above, the relation *m is the Gödel number of a proof in Q+ of the sentence the Gödel number of which is n*, as well as its negation, are representable in Q+. So, for any such Q+, we can identify certain theorems (which it asserts to be true) that "tell us" that certain numbers are Gödel numbers of provable sentences. When S isn't provable in Q+, a class of theorems will "tell us" for each natural number that it is *not* the Gödel number of a proof of that sentence.

This recapitulates *within Q+* much of the reasoning used to establish the simple version of the incompleteness result. However, there is a complication. Although the set of Gödel numbers of theorems of Q+ is *definable* in LA by the predicate $\exists x\, Proof_{Q+}\, x,y$, that set is not decidable (because there is no upper bound on how long we have to search for a proof to determine that S isn't provable). Thus, there is no guarantee that the set of theorems of an arbitrary extension Q+ of Q is *representable* in Q+. For example, $\exists x\, Proof_{Q+}\, x,y$ will *fail* to represent it, even though $Proof_{Q+}\, x,y$ *does* represent proof in Q+, if for some S that is not provable in Q+, $\sim\exists x\, Proof_{Q+}\, x,s$ is not provable in Q+, or, worse, $\exists x\, Proof_{Q+}\, x,s$ is provable in Q+, even though for each n, $\sim Proof_{Q+}\, \underline{n},s$ is provable in Q+. Each of these can happen in some consistent extensions of Q+. Since Gödelian reasoning reaches those systems too, we need a different proof.

3.2. *Omega-Consistent First-Order Extensions of Q Are Incomplete*

We begin with Gödel's (1931) statement of the theorem plus further needed definitions.

Gödel's Original Theorem
: All ω-consistent (i.e., omega-consistent) axiomatizable extensions of Q are incomplete.

Completeness
: A theory T is complete iff for every sentence S, either S or ⌜~S⌝ is a theorem of T.

Omega Completeness (ω-completeness)
: A theory T is ω-complete iff for all one-place predicates $F(\ldots x \ldots)$ in LA and all natural numbers n, if $F(\ldots \underline{n} \ldots)$ is a theorem of T, then so is $\forall x\, (\ldots x \ldots)$.

Consistency

A theory T is consistent iff no sentence and its negation are both theorems of T.

Omega-Consistency

A theory T is ω-consistent iff for all one-place predicates *F(. . . x . . .)* in LA and all natural numbers n, if *F(. . . n. . .) is* a theorem of T, then ~∀*x (. . . x . . .)* is *not* a theorem of T.

There are several points to note about these definitions. First, if Q+ is ω-consistent, then Q+ is consistent, but not conversely. Q+ will have a model that includes all natural numbers plus other things too, which means that Q+ *might* prove, for each number, that some predicate holds of it while also proving it fails to hold of something or other. Second, Q+ can be ω-complete without being complete, but not conversely. Third, any theory with N as a model is ω-consistent, but not conversely. Fourth, as we will see from Gödel's theorem, if Q+ is consistent, then it can't be ω-complete.

To prove the theorem, we need a consequence of a key Gödelian lemma establishing the existence of a certain sentence G about which Q+ *proves* (5) as a theorem, thereby making the paradoxical assertion that G is true iff there is no proof in Q+ of G.

5. G ↔ ~Prov G (where 'G' is replaced by the numeral designating the Gödel number of G)

The general statement of the lemma is (6).

6. If Q+ is a consistent extension of Q, then for every one-place predicate *B(y)* of LA, there is a sentence S such that Q+ ⊢ S ↔ B(s)

We begin with the one-place predicate (7) and its self-ascription G.[17]

7. ∃y (Self-Ascription x,y & B(y))

G ∃y (Self-Ascription $\underline{7}$, y & B(y))

Since self-ascription is representable in Q+, we have both (8a) and (8b), which gives us (9).[18]

8a. Q+ ⊢ Self-Ascription $\underline{7}$, G

8b. Q+ ⊢ G ↔ (Self-Ascription $\underline{7}$, G & B(G))

9. Q+ ⊢ G ↔ B(G)

Letting *~Prov y* be our choice for *B(y)*, gives us what we need to prove Gödel's theorem.

10. Q+ ⊢ G ↔ ~Prov G

[17] We let '$\underline{7}$' stand in for the numeral denoting the Gödel number of (7).

[18] Since the left side of (8b) is just the existential generalization of the right side, the right-to-left direction of (8b) is trivial. The left-to-right direction follows from the fact that the right side of (8b) is a logical consequence of G plus Q+ ⊢ ∀x (Self-Ascription $\underline{7}$, x → x = G).

This brings us to the proof of the original theorem. We let Q+ be an omega-consistent extension of Q. Our crucial sentence G is

G. ∃y (Self-Ascription κ, y & ~Prov y)

where *~Prov y* is the one-place predicate ~∃*x Proof*$_{Q+}$ *x,y*, k is its Gödel number, and *Proof*$_{Q+}$ *x,y* represents the set of Gödel numbers of proofs in Q+. Our key lemma, (10), tells us that Q+ proves (asserts): G iff Q+ G isn't provable—i.e., $G \leftrightarrow$ *~Prov* G. Next suppose, for *reductio*, that Q+ proves G. Since Q+ proves every logical consequence of things it proves, Q+ proves *~Prov* G. Since *Proof*$_{Q+}$*x,y* represents proof, there is a number m such that Q+ proves *Proof*$_{Q+}$*m*,G. Noting again that Q+ proves every logical consequence of anything it proves, we see it also proves *Prov* G. This contradicts our earlier result. *Thus, we establish that Q+ doesn't prove G.* Next we observe that no number m is the Gödel number of a proof of G, which, by the representability of proof, means that, *for each such m*, Q+ proves *~Proof*$_{Q+}$ *m, G*. Since Q+ is omega-consistent, Q+ *doesn't* prove ~∀x *~Proof*$_{Q+}$ *x, G*, or ∃*x Proof*$_{Q+}$ *x, G*—i.e., it *doesn't* prove *Prov* G. By (10) this means that Q+ doesn't prove *~G*. So, if Q+ is an ω-consistent, axiomatizable extension of Q, it proves neither *G* nor *~G*, hence it is incomplete.

It is striking that this proof relies on ω-*consistency*. Why wasn't simple consistency enough? Two factors conspired to force the narrower result. The first was reliance on *representability* rather than (semantic) *definability*. Here, there was no choice. Given the goal of proving the widest possible incompleteness result, relying on truth in N would have been too narrow. But the other factor—selecting the sentence G to use in applying the lemma $G \leftrightarrow B(G)$—was a matter of choice. By taking G to be ∃*y* (*Self-Ascription κ, y & ~Prov y),* we preserve a satisfying parallel with the simple semantic version of the theorem, at the price of forcing the proof to rely on ω-completeness. As we will see in the next section, that price is unnecessarily high.

First, however, I recapitulate the steps of the proof of the original formulation of the theorem.

PART 1

S1. Q+ is a *consistent* axiomatizable extension of Q.
S2. Then *Proof x,y* represents the set of pairs m,n such that m is the Gödel number of a proof in Q+ of the sentence the Gödel number of which is n.
S3. Letting G be ∃*y(Self-Ascription κ,y & ~Prov y),* we have Q+ ⊢ G ↔ ~Prov G (from our lemma).

PART 2

S4. Suppose Q+ ⊢ G
S5. Then some number m is the Gödel number of a proof in Q+ of G.
S6. From S2 and S5 we have Q+ ⊢ Proof m, G.
S7. From S6 we have Q+ ⊢ ∃x Proof x, G—i.e., Q+ ⊢ Prov G.
S8. From S3 and S7 we have Q+ ⊢ ~Prov G.

S9. Since Q_+ is consistent, S4 has led to a contradiction. So we have established *not* $Q_+ \vdash G$.

PART 3

S10. It follows from S9 that no natural number is the Gödel number of a proof in Q_+ of G.
S11. So, $Q_+ \vdash$ ~Proof $\underline{m}$, G for each natural number m.
S12. If, in addition to being consistent, Q_+ is ω-*consistent*, then we have:
Not $Q_+ \vdash$ ~∀x ~Proof x, G
Not $Q_+ \vdash$ ∃x Proof x, G—i.e., Not $Q_+ \vdash$ Prov G
S13. Suppose $Q_+ \vdash$ ~G.
S14. Then, from S3 and S13, we have $Q_+ \vdash$ Prov G.
S15. Since S14 contradicts S12, S13 is false.
S16. So, if Q_+ is ω-consistent, then neither $Q_+ \vdash G$ nor $Q_+ \vdash$ ~G.

3.3. The Rosser Extension: Consistent, Axiomatizable First-Order Extensions of Q Are Incomplete

The strengthened Gödel-Rosser theorem states that all consistent, axiomatizable first-order extensions of Q are incomplete.[19] It differs from Gödel's original theorem in its choice of G. In the original proof, we used a sentence G that says *I am not provable*—in the sense of "say" in which Q_+ asserts (proves) *that G is true iff the claim that G is not provable is true*. In the Rosser proof, we use a sentence G* that "says" that *if I am provable, then some number that encodes a proof of my negation is smaller than any number that encodes a proof of me*—in the sense of "say" in which Q_+ asserts (proves) *that G* is true iff, if G* is provable, then some number that encodes a proof of the negation of G* is smaller than any number that encodes a proof of G**.

To express Rosser's new "Gödel sentence" G* (in the metalanguage), we introduce a one-place function symbol *Neg* that represents the computable function that maps the Gödel number of a one-place predicate of the language of arithmetic onto the Gödel number of its negation (and everything else onto zero). We then consider the one-place (metalanguage) predicate (7*) that represents the set of Gödel numbers of sentences S that "say" *if S is provable, then there is a proof of S's negation encoded by a smaller number than any number that encodes a proof of S.*

7*. [Self-Ascription *x,y* & Prov y → (∃x (Proof x, Neg(y) & ∀z (Proof z, y → x < z)))]

Taking '$\underline{7^*}$' to denote the Gödel number of (7*), we call the self-ascription of (7*) "G*".

[19] Rosser (1937).

G*. ∃y [Self-Ascription 7*, y & Prov y → (∃x (Proof x, Neg(y) & ∀z (Proof z, y → x < z))]

The instance of the lemma (9) established above is 9*, where ~G* is the numeral denoting the Gödel number of the negation of G*.

9*. $Q_+ \vdash G^* \leftrightarrow [\text{Prov } \ulcorner G^* \urcorner \rightarrow \exists x (\text{Proof } x, \ulcorner \sim G^* \urcorner \,\&\, \forall z (\text{Proof } z, \ulcorner G^* \urcorner \rightarrow x < z))]$

Although the crucial Gödel sentence G* and lemma 9* are more complex than the originals, the reason for the extra complexity is easy to see. In the original proof, the fact that G isn't a theorem follows directly from lemma 9 and the consistency of the system. Omega-consistency is needed to show ~*G* isn't a theorem. We know, for each natural number n, ~*Proof* $\underline{n}$, $\ulcorner G \urcorner$ is a theorem, but we need to rule out the possibility that, *for each model M of* Q_+, there is something o* in the domain of M (other than a natural number) such that *Proof* is true of < o*, G >. Q_+ is omega-consistent, and hence that ∃*x Proof x,*$\ulcorner G \urcorner$ (~∀*x* ~*Proof x,* $\ulcorner G \urcorner$) isn't a theorem. This plus the supposition that ~*G* is a theorem (and lemma 9) gives us a contradiction, showing ~*G* isn't a theorem.

Rosser's selection of G* circumvents the need for omega-consistency by allowing us to derive, from the assumption that ~*G** is a theorem, (i) that for some number m, *Proof* $\underline{m}$, $\ulcorner \sim G^* \urcorner$ is a theorem, and that, on pain of inconsistency, for each number i ≤ m, *Proof i,* $\ulcorner G^* \urcorner$ isn't a theorem, while ∀*x (Proof x,* $\ulcorner G^* \urcorner \rightarrow \underline{m} < x$) is a theorem. Lemma 9* then gives us the result, (ii) that ~[∃*y Proof y* $\ulcorner G^* \urcorner \rightarrow \exists x$ *(Proof x,* $\ulcorner \sim G^* \urcorner$ *&* ∀*z (Proof z,* $\ulcorner G^* \urcorner \rightarrow x < z)$)] is a theorem, from which we derive that ∃*y (Proof y,* $\ulcorner G^* \urcorner$ *&* ~*m < y)* is too. Since this contradicts (i), we conclude that ~*G** isn't a theorem.

S1. Suppose $Q_+ \vdash \sim G^*$
S2. Then, some number m is the Gödel number of a proof in Q_+ of ~*G**.
S3. $Q_+ \vdash$ Proof $\underline{m}$, $\ulcorner \sim G^* \urcorner$ (since *Proof x,y* represents proof in Q_+)
S4. By consistency of Q_+, for all numbers i ≤ m, i is *not* the Gödel number of a proof of *G**.
S5. For each number i < m, $Q_+ \vdash$ ~Proof $\underline{i}$, $\ulcorner G^* \urcorner$ (since *Proof x,y* represents proof in Q_+)
S6. $Q_+ \vdash \forall x (x < \underline{m} \rightarrow x = \underline{0} \vee x = \underline{1} \vee \ldots \vee x = \underline{m} - \underline{1})$ (Provable in Q_+)
S7. $Q_+ \vdash \forall x (x < \underline{m} \vee x = \underline{m}) \rightarrow \sim$Proof x, $\ulcorner G^* \urcorner$) (from S2, S5, S6)
S8. $Q_+ \vdash \forall x$ (Proof x, $\ulcorner G^* \urcorner \rightarrow \sim(x < \underline{m} \vee x = \underline{m})$) (from S7)
S9. $Q_+ \vdash \forall x (x < \underline{m} \vee x = \underline{m} \vee \underline{m} < x)$ (Provable in Q)
S10. $Q_+ \vdash \forall x$ (Proof x, $\ulcorner G^* \urcorner \rightarrow \underline{m} < x$) (from S8, S9)
S11. $Q_+ \vdash \sim[\exists y$ Proof y $\ulcorner G^* \urcorner \rightarrow \exists x$ (Proof x, $\ulcorner \sim G^* \urcorner$ & $\forall z$ (Proof z, $\ulcorner G^* \urcorner \rightarrow x < z))]$ (from S1 and lemma 9*)
S12. $Q_+ \vdash \exists y$ Proof y $\ulcorner G^* \urcorner$ & $\forall x \sim$(Proof x, $\ulcorner \sim G^* \urcorner$ & $\forall z$ (Proof z, $\ulcorner G^* \urcorner \rightarrow x < z)$) (from S11)
S13. $Q_+ \vdash \exists z$ (Proof y, $\ulcorner G^* \urcorner$ & ~$\forall z$ (Proof z, $\ulcorner G^* \urcorner \rightarrow \underline{m} < z$) (from S3 and S12)
S14. $Q_+ \vdash \exists z$ (Proof z, $\ulcorner G^* \urcorner$ & ~$\underline{m} < z$) (from S13)
S15. Since this contradicts S10, the consistency of Q_+ requires S1 to be rejected; ~*G** is *not* a theorem of Q_+.

W show that G* isn't a theorem of Q+ in the same way. Starting with the supposition that *G** is a theorem, we show (i) that for some number m *Proof* $\underline{m}$, *G** is a theorem, and that, on pain of inconsistency, for each number i ≤ m, *Proof* $\underline{i}$, ~*G** isn't a theorem, while ∀x (Proof x, ~G* → ~x < $\underline{m}$) is a theorem. Adding lemma 9* gives us (ii) the theoremhood of *[Prov G* → ∃x (Proof x, ~G* & ∀z (Proof z, G* → x < z))],* which allows us to derive ∃*x (Proof x, ~G* & x <* $\underline{m}$*))],* which contradicts the result of (i). Thus, we conclude that G* isn't a theorem; the consistency of Q+ is sufficient to guarantee that it is incomplete.

S1. Suppose Q+ ⊢ G*.
S2. Then, some number m is the Gödel number of a proof in Q+ of G*.
S3. Q+ ⊢ Proof $\underline{m}$, G* (since *Proof x,y* represents proof in Q+).
S4. By consistency of Q+, for all numbers i < m, i is *not* the Gödel number of a proof of ~*G** (since otherwise ~*G** would be provable, since *Proof x,y* represents proof in Q+).
S5. For each number i < m, Q+ ⊢ ~Proof $\underline{i}$, ~G* (since *Proof x,y* represents proof in Q+).
S6 Q+ ⊢ ∀x (Proof x, ~G* → ~x < $\underline{m}$) (from S5) .
S7. Q+ ⊢ [Prov G* → ∃x (Proof x, ~G* & ∀z (Proof z, G* → x < z))] (from S1 and lemma 9*).
S8. Q+ ⊢ ∃x (Proof x, ~G* & x < $\underline{m}$) (from S7, S3) .
S9. Since this contradicts S6, the consistency of Q+ requires S1 to be rejected; *G** is *not* a theorem of Q+.

3.4. Non-Categoricity and Categorical Second-Order Arithmetic

In showing that all consistent first-order axiomatizable extensions Q+ of Q are incomplete, we have also shown that they are all ω-*incomplete*. Recall the original version of the theorem in which the relevant unprovable sentence is G, which is a self-ascription of (1), and the corresponding instance of our lemma is (10).

1. ∃y (Self-Ascription x,y & ~Prov y)
G. ∃y (Self-Ascription $\underline{1}$, y & ~Prov y)
10. Q+ ⊢ G ↔ ~Prov G

We saw that G is not provable in Q+. Thus for each natural number n, ~*Proof* $\underline{n}$, *G** is a theorem of Q+. If Q+ were ω-complete, ∀x ~*Proof x,G*, ~*Prov G*, and (by (10)), G would be theorems of Q+. Since this can't be, Q+ is not ω-complete.

Next consider all consistent first-order axiomatizable extensions Q+ of Q that are true in the intended model N (the domain of which includes all and only the natural numbers). Since the set of true sentences of LA always includes ∀*xΦx* whenever it includes *Φ*$\underline{n}$ for each natural number n, every Q+ that has N as a model must have nonstandard models the domains of which are *not* isomorphic with the set of natural numbers ordered under

successor. In other words, the structure of the natural numbers cannot be identified up to isomorphism by any axiomatizable first-order theory.[20]

In itself, this is not surprising. It is a theorem of the metatheory of first-order logic that any set of first-order sentences that has infinite models has both models with countable domains and models with uncountable domains.[21] Since models with different cardinalities can't be isomorphic, this means that there is no consistent set of first-order sentences all models of which are isomorphic. There *are* consistent sets of first-order sentences of LA with countably infinite models all of whose countable models are isomorphic.[22] However, none are complete. Thus, it is a corollary of the first incompleteness theorem that there are no consistent extensions of Q all of whose *countable* models are isomorphic.[23]

It may be more surprising to learn that the set T_A of sentences of LA that are true in the intended model N also has models with *countable* domains that are not isomorphic with the intended model N. Though this result wasn't proved by Gödel, it does contribute to our understanding of why his first incompleteness theorem applies to first-order axiomatizations of arithmetic, but not to second-order axiomatizations. The domain of a countable, nonstandard model of T_A will contain an *initial* segment ordered under *less than** that is identical to, or isomorphic with, the natural numbers (ordered under *less than*). This segment will be followed by *blocks* of linearly ordered elements each of which is isomorphic with the series of all integers (negative, zero, and positive). Each element of one of these blocks will have a unique *successor** (and *predecessor**)—where *successor** (*predecessor**), addition*, multiplication*, and *less than** operate throughout the domain of the nonstandard model, just as *successor* (*predecessor*) *addition, multiplication,* and *less than* do throughout the domain of the intended model of first-order arithmetic. Although all elements, standard and nonstandard, in the model are linearly ordered with respect to one another, there is no *least* and no *greatest* block of nonstandard elements. Also, between any two such blocks there is another block. Since there are countably many such blocks, the nonstandard domain is itself countable. Clearly a nonstandard model of T_A is *not* isomorphic with the intended model N of T_A, even though the two models assign truth (and falsity) to the same sentences of LA.[24]

If we want an axiomatizable arithmetical theory all models of which are isomorphic with the intended model N—and thus a theory that has each

[20] A set of sentences is *categorical* iff all its models are isomorphic.
[21] Lowenheim (1915), Skolem (1920).
[22] Such sets are said to be *aleph-null categorical.*
[23] The proof is trivial. See Boolos and Jeffrey (1974), p. 193.
[24] The proof of this result, presented in Boolos and Jeffrey (1974), pp. 194–96, is too detailed to detain us here. Though not challenging, the technique is interesting and well worth looking at.

(first-order) arithmetical truth as a logical consequence—we must include a second-order sentence among the axioms. We can do this by replacing the infinitely many instances of the induction axiom schema of first-order Peano arithmetic with the induction axiom of second-order Peano arithmetic.

THE FIRST-ORDER AXIOM SCHEMA OF INDUCTION

$$[(F(0) \,\&\, \forall x\,(F(x) \rightarrow F(S(x)))) \rightarrow \forall x\, F(x)]$$

Instances are formed by replacing 'Fx' with one-place predicates of LA, and 'F(0)' with the result of replacing free occurrences of 'x' in the formula that replaces 'Fx' with occurrences of '0'

THE INDUCTION AXIOM OF SECOND-ORDER PA

$$\forall P\,[(P(0) \,\&\, \forall x\,(P(x) \rightarrow P\,(S(x)))) \rightarrow \forall x\, P(x)]$$

To interpret this axiom we need to add to the definition of *truth in a model* given in section 1.2 above. Nothing changes in the conception of a model that interprets a theory's nonlogical symbols over a domain that is the range of the first-order quantifiers. We simply add assignments of subsets of the domain to one-place second-order predicate variables (plus sets of n-tuples of individuals of the domain to n-place second-order predicate variables, etc.). The truth of a formula is now relativized to an assignment of values to first-order variables plus an assignment of values to second-order predicate variables. Where D is the domain of M, the extra clauses in the definition of *truth in a model* are:

- $\ulcorner \exists P(\ldots P \ldots) \urcorner$ is true in M relative to a pair of assignments A_1 of individuals in D and A_2 of sets of n-tuples of D to n-place second-order (predicate) variables iff there is at least one such set D_S of n-tuples for which $(\ldots P \ldots)$ is true in M relative to A_1 and an assignment $A_2{}^*$ that differs at most from $A_2{}^*$ in assigning D_S to P.[25]
- $\ulcorner \forall P(\ldots P \ldots) \urcorner$ is true in M relative to a pair of assignments A_1 of individuals in D and A_2 of sets of n-tuples of D to n-place second-order (predicate) variables iff for every such set D_S of n-tuples, $(\ldots P \ldots)$ is true in M relative to A_1 and an assignment $A_2{}^*$ that differs at most from $A_2{}^*$ in assigning D_S to P.

With this in mind, let M be any model of second-order PA. Since D is a subset of itself, the truth of the induction axiom in M tells us its domain is isomorphic with the set of natural numbers ordered under successor. So, second-order PA has no nonstandard models.[26] Since M is isomorphic

[25] $(\ldots P \ldots)$ arises from the quantified formula by erasing the quantifier. An atomic formula consisting simply of an n-place predicate variable P plus n terms is true relative to assignment A of individuals to first-order variables plus an assignment of a set D_S of members of D to P iff the n-tuple of referents of the terms relative to A is a member of D_S.

[26] See Boolos and Jeffrey (1974), pp. 203–4.

with the intended model N of first-order arithmetic, the first-order sentences true in M are all and only the first-order arithmetical truths. (The logical consequences of second-order PA are all and only the first- or second-order truths of arithmetic.) Thus, second-order PA is *complete in the sense of Gödel's first incompleteness theorem*.[27]

However, this doesn't subvert the significance of that theorem. Although all first-order truths of arithmetic are *logical consequences* of second-order Peano arithmetic, second-order logic is *not complete in the sense in which first-order logic was proved complete in Gödel (1930)*. In first-order logic *every logical truth* is *provable* from a consistent set of logical axioms and rules of inference—where a *proof* is a finite sequence of lines, each of which is an axiom or a formula obtainable from earlier lines by the inference rules. Similarly, *every logical consequence* B of a first-order sentence A (or of a decidable set A* of first-order sentences) is *provable from A* (or from a finite subset of A*). By contrast, second-order logic is *not complete* in this sense. For any consistent system of proof for second-order logic, there will be second-order logical truths that are not provable in the system, and there will be logical consequences of second-order sentences (or of decidable sets of such) that are not provable from those sentences (or sets) in the system. So, although second-order Peano arithmetic is a *complete* formal theory that, for each arithmetical sentence S, has S or ⌜~S⌝ as a *logical consequence*, the fact that the *logic* of second-order consequence is *incomplete*—in the sense that the *logic* of first-order consequence is complete—means that there is no effective positive test for first-order arithmetical truth. Thus, the chief lesson of Gödel's first incompleteness proof is untouched.

4. GÖDEL'S SECOND INCOMPLETENESS THEOREM: THE UNPROVABILITY OF CONSISTENCY

The second incompleteness theorem is an elaboration of the first. Having seen that consistent first-order axiomatizations of arithmetic must fail to prove infinitely many arithmetical truths, one naturally wonders, *What more can be said about the scope and nature of the truths that remain outside the range of any such system?* The first incompleteness theorem was based on a sentence G that "says of itself" that it is not provable in a given consistent system S. What about a sentence that "says," *If I am provable in S, then I am true,* or better, one that "says" of S that *no contradiction is provable in S* (and hence that *S is consistent*)? The aim of the second incompleteness theorem is to show that although sentences "asserting the consistency" of PA can be identified, no such sentence is a theorem of PA.[28] Our proof will also

[27] Ibid., Corollary 1 on p. 204.

[28] As before, the use of Q to specify the scope of the theorem occurred years after the original proof.

answer the question about the first sort of sentence—which "says" it is true if provable in PA.

Torkel Franzen provides the following informative historical commentary on the origin of the second incompleteness theorem.

> Gödel first presented his incompleteness theorem at a conference on "Epistemology of the exact sciences" in 1930. . . . Among those present was the Hungarian mathematician John von Neumann. . . . It appears he was the one participant at the conference who immediately understood Gödel's proof. At this point Gödel had not yet arrived at his second incompleteness theorem, and his proof of the first incompleteness theorem was not applicable to PA, but only to somewhat stronger theories. His proof did, however, establish that assuming a theory S . . . to be consistent, it follows that the Gödel sentence G for S is unprovable in S. Reflecting on this after the conference, von Neumann realized that the argument establishing the implication "If S is consistent, then G is not provable in S" can be carried out within S itself. But then, since G is equivalent in S to "G is not provable in S" [which is an instance of our lemma above], it follows that if S proves the statement Con_S expressing "S is consistent" in the language of S, S proves G, and hence is in fact inconsistent [because then both *Prov* $\ulcorner G \urcorner$ and ~*Prov* $\ulcorner G \urcorner$ will be theorems]. Thus, the second incompleteness theorem follows: If S is consistent, Con_S is not provable in S. By the time von Neumann had discovered this and written to Gödel about it, Gödel himself had already made the same discovery and included it in his recently accepted 1931 paper.[29]

As Franzen points out, Gödel merely sketched the idea behind the second incompleteness theorem, without presenting a detailed proof.[30] Although a rigorous proof didn't appear until Hilbert and Bernays (1939), the theorem was accepted as a corollary of the first incompleteness theorem on the informal grounds offered in Gödel (1932). We know from the first incompleteness theorem that if PA is consistent, then G is not provable in PA, but $G \leftrightarrow$ ~*Prov* $\ulcorner G \urcorner$ is. If, as maintained in Gödel (1932), this reasoning is expressible in PA, then both $Con_S \rightarrow$ ~*Prov* $\ulcorner G \urcorner$ and $G \leftrightarrow$ ~*Prov* $\ulcorner G \urcorner$ are provable in PA. If Con_S were also provable, both ~*Prov* $\ulcorner G \urcorner$ and G would be theorems. Since this would mean that *Prov* $\ulcorner G \urcorner$ was also a theorem, PA would be inconsistent. Hence, Con_S must *not* be provable in PA.

That is the informal idea behind the second incompleteness theorem. In the decades that followed, various fully detailed proofs of the theorem were discovered. I will sketch one based on Lob's theorem given in Lob (1955). We begin with two questions.

> Q1. We have seen that there are sentences G that can be taken as truly asserting that "G is not provable in PA." Are there sentences H that can be

[29] Franzen (2005), pp. 97–98.
[30] Ibid., p. 98.

taken as asserting "H is provable in PA," and, if so, are they, or are they not, provable in PA?

Q2. Are there sentences of LA that can be taken as asserting "the consistency of PA," and, if so, are they, or are they not, provable in PA?

We can answer the first part of these questions right away. By the Gödelian lemma (6),

6. If Q+ is a consistent extension of Q, then for every one-place predicate *B(y)* of LA, there is a sentence S such that $Q+ \vdash S \leftrightarrow B(s)$

We have a sentence H such that $PA \vdash H \leftrightarrow$ Prov H. Since *Prov H* just is $\exists x$ *Proof x, H*—while *Proof x,y* defines (represents) proof in PA—it is demonstrable that H is true (in the intended interpretation N) iff H is provable in PA. Hence, H can be taken as "saying" that H is provable in PA. Next, PA is consistent iff it proves no contradictions. Since PA is an extension of Q, it proves $0 \neq 1$. So if PA is consistent, $0 = 1$ isn't a theorem, and, if $0 = 1$ isn't a theorem, PA is consistent (since inconsistent theories prove every sentence). Thus, by the reasoning we just went through with H, *~Prov 0=1* can be taken to "say" that PA is consistent.

Are H and *~Prov 0=1* provable in PA? Lob's theorem will show us that H is provable, but *~Prov 0=1* isn't, thus establishing the second incompleteness theorem. To simplify matters, we assume that PA (and other systems under consideration) employ only one logical rule of inference, *modus ponens*—which allows one to infer *C* from *A* and $A \rightarrow C$. With this in mind, we note that our predicate *Prov y* for PA satisfies conditions C1–C4 for any sentence A of LA.[31]

C1. If $PA \vdash A$, then $PA \vdash$ Prov A
C2. $PA \vdash (\text{Prov } A \rightarrow C \rightarrow [\text{Prov } A \rightarrow \text{Prov } C])$
C3. If $PA \vdash A$, then $PA \vdash (\text{Prov}_{\text{Prov } A})$
C4. If $PA \vdash$ Prov A, then $PA \vdash A$

C1 and C4 follow directly from the fact that *Proof x,y* defines (represents) proof in PA and *Prov y* is the formula $\exists x$ *Proof x, y*.[32] C2 results from the fact that when $(A \rightarrow C)$ and *A* are both provable in PA, one can always get a proof of *C* by combining the two proofs, one after the other, and adding *C* as the last line. C3 tells us that if PA proves that *A is provable*, then it proves that *it proves that it proves that A is provable*. Although this is harder to establish, it can also be shown.[33]

[31] To get C3, first substitute the numeral denoting the Gödel number of A for the free variable *y* in the one-place predicate *Prov y*. Let k be the Gödel number of the resulting sentence. Then substitute the numeral denoting k for *y* in *Prov y*. The result is the consequent of the theorem mentioned in C3.

[32] N is a model of PA. The induction axioms of PA guarantee that $\exists x$ *Proof x,* $\underline{A}$ can be a theorem only if *Proof* $\underline{n}$, $\underline{A}$ is a theorem for some natural number, in which case A is a theorem.

[33] See Hilbert and Bernays (1939) and Lob (1955).

Next we define Lob's notion of a *provability predicate for a theory T*, where the only constraint we impose on T is that it be an extension of Q.

> If T is an extension of Q, then a formula *B(y)* of LA in which the variable *y* is the only variable with free occurrences is a *provability predicate for T* iff for all sentences A and C of the language of T,
>
> C1. If $T \vdash A$, then $T \vdash B_A$
>
> C2. $T \vdash (B_{(A \to C)} \to [B_A \to B_C])$
>
> C3. $T \vdash (B_A \to B_{B(A)})$

We have seen that *Prov y* is a *provability predicate* for PA. It is worth noting, however, that despite the name "provability predicate," the class of predicates so defined includes some that don't have anything to do with proof. For example, if T is a consistent extension of Q and the one-place predicate *S(y)* represents the set of Gödel numbers of sentences of LA, then *S(y)* qualifies as a *provability predicate* for T. C1 is satisfied, since for any theorem A of T there will be another that can be taken to "say" that A is a sentence of LA. C2 is satisfied, since for any sentences A, C of LA, among the sentences provable in T will be one that can be taken to "say" that if ⌜A → C⌝ is a sentence, then if A is a sentence so is C. C3 is satisfied because, for any sentence A of LA, there is a sentence S1 of LA that can be taken to "say" that A is a sentence of LA and there is another sentence S2 of LA that can be taken to "say" that S1 is a sentence of LA, and the conditional with S1 as antecedent and S2 as consequent will be a theorem of T. In short, the formal notion *provability predicate* defined here captures some essential features of predicates that genuinely encode *proof*. But it's not crucial for Lob's theorem that this notion capture all their essential features.

Lob's theorem uses the notion just defined. It says:

> If *B(y)* is a provability predicate for an extension T of Q, then for any sentence A of the language of T, if $T \vdash (B_A \to A)$ then $T \vdash A$.

Before proving the theorem, it is worth noting its relation to Gödel's incompleteness theorems. The first point concerns the extent of the incompleteness established by Gödel's first theorem. Let Q+ be any consistent extension of Q that proves only sentences true in the intended model N. Since *Prov y* is true of only Gödel numbers of truths of arithmetic, *(Prov A → A) will be true for every sentence A (true or false).* Since *Prov y* is a provability predicate for Q+, Lob's theorem indicates how massively incomplete these theories are.

The second point concerns our earlier question about whether a sentence H that can be taken to assert its own provability is itself provable in PA. Since PA is an extension of Q and *Prov y* is a *provability predicate* for PA, Lob's theorem tells us that if *(Prov H → H)* is provable in PA, then H is too. Since we already know that $H \leftrightarrow$ *Prov H* is provable in PA (from our lemma 6 that was crucial to the first incompleteness theorem) we see that H *is* provable in PA. Like G, which asserts its own *unprovability*, H, which asserts its own *provability*, is true (in N).

Finally, consider the sentence *~Prov 0=1,* which, as we have seen, can be taken to "say" that PA (or any other consistent extension of Q) is consistent. Suppose, for *reductio,* that it is a theorem of PA (or any other consistent extension of Q). Then both *(~Prov 0=1 ∨ 0 = 1)* and *(Prov 0=1 → 0 = 1)* will also be theorems. From Lob's theorem plus the fact that *Prov y* is a provability predicate for PA (or for any consistent Q_+), we get the result that *0 = 1* is also a theorem—*which it can't be,* on the assumption that our formal system *is* consistent. So, *~Prov 0=1* is not a theorem. Thus, if Q_+ is consistent, no statement "asserting" its consistency can be proven in Q_+. This is Gödel's second incompleteness theorem: *consistent, axiomatizable first-order theories of arithmetic are incapable of proving their own consistency*.

All that remains is to prove Lob's theorem.

Prove: If *B(y)* is a provability predicate for a consistent extension T of Q, then for any sentence A, if $T \vdash (BA \rightarrow A)$ then $T \vdash A$.

S1. Assume $T \vdash (BA \rightarrow A)$
S2. Since *B(y)* → *A* is a one-place predicate of LA, our earlier lemma—6 above—tells us that there is a sentence C such that $T \vdash C \leftrightarrow (BC \rightarrow A)$
S3. $T \vdash C \rightarrow (BC \rightarrow A)$ From S2
S4. $T \vdash B(C \rightarrow (B(C) \rightarrow A))$ From S3 and C1 of the definition of *provability predicate*
S5. $T \vdash B[(C \rightarrow (B(C) \rightarrow A)) \rightarrow [BC \rightarrow B(B(C \rightarrow A))]]$ Instance of C2 of the definition of *provability predicate*
S6. $T \vdash BC \rightarrow B(B(C \rightarrow A))]$ From S4, S5
S7. $T \vdash B(B(C \rightarrow A)) \rightarrow [B(B(C) \rightarrow B(A)]$ Instance of C2 of the definition of *provability predicate*
S8. $T \vdash BC \rightarrow [B(B(C) \rightarrow B(A)]$ From S6, S7
S9. $T \vdash BC \rightarrow B(B(C))$ Instance of C3 of the definition of *provability predicate*
S10. $T \vdash BC \rightarrow BA$ From S8, S9
S11. $T \vdash BC \rightarrow A$ From S1, S10
S12. $T \vdash C$ From S2, S11
S13. $T \vdash BC$ From S12 and C1 of the definition of *provability predicate*
S14. $T \vdash A$ From S11, S13
S15. $T \vdash (BA \rightarrow A)$ then $T \vdash A$ From S1, S14

5. COMPUTABILITY AND UNDECIDABILITY

One of the unmistakable lessons of Gödel's theorems is the intimate connection between *effective computability* and *proof* in an axiomatized formal system. *Proof* in a formal system of the sort to which the theorems apply is always an effectively decidable notion. Moreover, each of these systems is rich enough to guarantee that systematic searches of proofs in it qualify as decision procedures for membership in decidable sets of (n-tuples of) natural numbers—and, by extension, for membership in any set decidably

coded by natural numbers. In light of this close relationship between computability and logic, it is not surprising that the notion of computability itself, and its further implications for logic, was intensely studied by philosophically minded logicians throughout the decade of the 1930s.

Chief among them was Alonzo Church, then Assistant Professor in the Princeton mathematics department.[34] His colleagues included Kurt Gödel, who visited in 1933–34 and again in 1935 before permanently moving to the Institute for Advanced Studies at Princeton at the outbreak of World War II. Church also supervised dissertations of a remarkably gifted group of his PhD students, including J. Barkley Rosser, Stephen Kleene, and Alan Turing. Church's chief concerns at the time were (i) formalizing, to the extent possible, the intuitive notion of an effectively computable function, and (ii) using that notion, in conjunction with Gödel's first incompleteness theorem, to prove that there can be no effective procedure for deciding whether an arbitrary first-order sentence is, or isn't, a logical truth (or a logical consequence of other first-order sentences). The latter, known as Church's Theorem, is proved in Church (1936a), which appeared in the first issue of the *Journal of Symbolic Logic*, of which Church was a founder and longtime editor.[35]

The former—the mathematical formalization of the intuitive notion of computability—was the result, at least in part, of a fruitful interchange with Gödel. In his 1934 lectures at Princeton, Gödel introduced recursive functions (definable by certain "recursive" equations) as a possible formalization. Church (1936b) introduced an alternative—functions definable by certain formulas of his λ-*calculus*. In addition to proving his notion to be equivalent to Gödel's, he hypothesized—in what is now known as *Church's* (so far unrefuted) *Thesis*—that it will be equivalent to any adequate formalization of the intuitive idea.[36] Turing was studying at Cambridge when Church's paper was published and had, himself, independently developed a third, very natural, formalization—which provided the basis for its own proof of the unsolvability of the decision problem for first-order logic.[37] After reading the paper, Turning wrote to Church, arranged a visit, and ended up, after encouragement by John von Neumann, remaining in Princeton to earn his PhD in 1938 under Church's supervision. The rest of this section will sketch these developments.

5.1. Church's Undecidability Theorem

As Church saw, his theorem is really a corollary of the first incompleteness theorem. Recall the so-called Gödel sentence G of the original proof that

[34] Church was promoted to Associate Professor in 1939.
[35] Church (1936a).
[36] Church (1936b).
[37] Turing (1936/37).

says that G is not provable. G is the self-ascription of the predicate (1) of LA (the Gödel number of which is designated by the numeral 1*) that is true of all and only the Gödel numbers of predicates of LA that are not provable of themselves. (See section 3.4.)

1. ∃y (Self-Ascription x,y & ~Prov y)
G. ∃y (Self-Ascription 1*, y & ~Prov y)

The predicate *Prov* y is the one-place predicate ∃*x Proof x,y*, which is an existential generalization of the two-place predicate *Proof x,y* that represents the decidable relation that holds between a pair of numbers iff the first codes a proof of the sentence that the second encodes. Evaluated at the intended model N, *Prov y* is true of a number iff it encodes a sentence provable in the system. One might think that this predicate *represents* the set of Gödel numbers of provable sentences. But a moment's thought will convince one otherwise. If it did, then the fact that G is, as Gödel showed, unprovable would guarantee that ~*Prov* g was provable, from which it would follow that G was provable after all. Since this can't be, ∃*x Proof x,y* doesn't represent the set of Gödel numbers of provable sentences. But then, one is inclined to think, no predicate does. This is the key to Church's Theorem.

The result we need is that *if* Q_+ is a consistent, axiomatizable first-order extension of Q, then the set TH of Gödel numbers of theorems of Q_+ is not representable in Q_+—i.e., there is no one-place predicate *Prov*y* of LA such that for any natural number n, if n is a member of TH, then $Q_+ \vdash$ Prov* $\underline{n}$, and if n isn't a member of TH, then $Q_+ \vdash$ ~Prov* $\underline{n}$.

S1. Suppose there is a formula *Prov** that represents TH in Q_+.
S2. As in the proof of the first incompleteness theorem, we have it that *for every one-place predicate B(y)* of LA, there is a sentence S such that $Q_+ \vdash S \leftrightarrow B(s)$.
S3. Consider the one-place predicate Z and its self-ascription G.[38]
    Z ∃y (Self-Ascription x,y & B(y))
    G ∃y (Self-Ascription z, y & B(y))
    Since self-ascription is representable in Q_+, we have
    $Q_+ \vdash$ Self-Ascription z, g
    $Q_+ \vdash G \leftrightarrow$ (Self-Ascription z, g & B(g))
    $Q_+ \vdash G \leftrightarrow B(\underline{g})$
    Letting ~*Prov*y* be our choice for *B(y)* in S2, we have
    $Q_+ \vdash G \leftrightarrow$ ~Prov* g
S4. Suppose not $Q_+ \vdash G$.
S5. From S1, S4 we have $Q_+ \vdash$ ~Prov* g.
S6. From S3, S5 we have $Q_+ \vdash G$.
S7. Since S6 contradicts S4, we have $Q_+ \vdash G$.
S8. S1 and S7 give us $Q_+ \vdash$ Prov* g.

[38] We let '$\underline{Z}$' stand in for the numeral denoting the Gödel number of (7).

S9. S3 and S7 give us Q+ ⊢ ~Prov* g.
S10. S8 and S9 contradict the consistency of Q+.
S11. Since the supposition that TH is representable in Q+ leads to contradiction, it follows that TH is not representable. The set of Gödel numbers of theorems is not representable in any consistent, first-order, axiomatizable extension of Q.

This all but proves Church's Theorem. For suppose that first-order logical consequence is decidable. Then the set of logical consequences of first-order, axiomatizable extensions Q+ of Q is decidable. By Gödel's completeness theorem for first-order logic (his dissertation), the logical consequences of Q+ are just its theorems—i.e., the sentences provable in Q+. So, if first-order logical consequence is decidable, the set of Gödel numbers of theorems of Q+ is representable in Q+. Since we have proved that it isn't, it follows that -order logic is undecidable.

5.2. Turing Machines, Turing-Computable Functions, and Halting Problem

The notion of a *Turing-computable function,* unlike its provably equivalent predecessors, *recursive function* and λ*-definable function,* is based on a transparently recognizable model of simple (deterministic) computing done either by a human agent or by a simple machine following elementary instructions. According to H. B. Enderton, the name "Turing machine" given to the central notion was first coined by Church in his review in the *Journal of Symbolic Logic* of Alan Turing (1936/37).[39] Although a Turing machine is a purely mathematical object—a finite set of quadruples—it is standardly thought of as a simple machine the instructions of which are identified with the quadruples.

The machine is imagined to operate on an infinite tape divided into squares, each of which is either blank or imprinted with a single dot. It can move along the tape, checking to see whether or not the square it is on is blank. It can also print a dot on a previously blank square or erase a dot on a square that had one. The machine has a finite number of internal states; its instructions tell it what to do, based on the state it is in at a given time. The first symbol Q_i in any instruction designates one of the machine's states. The second symbol is either S_0, to signal scanning a blank, or S_1, to signal scanning a square with a dot.[40] The third symbol of the instruction tells the machine what to do *if it is in the state designed by the first symbol of the instruction scanning a square of the type indicated by the second symbol.* If it is

[39] Enderton (1998), Church (1937).
[40] Although the number of symbols a machine is capable of recognizing (and printing) can be increased to any finite number, this has no effect on the functions computable by Turing machines. Zero and one are enough.

S_1 the machine prints a dot on a previously blank square; if it is S_0 the machine erases the dot it is scanning; if it is *L* the machine moves one square to the left; if it is *R* the machine moves one square to the right. The final symbol in an instruction specifies the internal state the machine is to be in (either the current state or a different one) after performing the action. Once started, a machine will continue until it reaches a state, scanning a square, which is not matched by an instruction telling it what to do.

What is it for a (numerical) function f to be computed by a Turing machine M? To answer this question, we must specify how the tape on which M is started is interpreted to represent the argument(s) of f, and how values of f can be read off the state of the tape when and if M halts after being started on that input. Although different conventions are possible, the following will suffice. The arguments of an n-place (numerical) function f are represented by n blocks of 1's (i.e., adjacent squares each with a dot), each block separated from the next by a single blank square, on an otherwise blank tape. Limiting ourselves to functions from positive integers to positive integers, we interpret a block of k 1's as representing the number k. M is always started scanning the leftmost 1 (a square with a dot) on the tape. If, after being started on a tape representing an argument, M eventually halts scanning the leftmost of a single block of 1's on an otherwise blank tape, then the number of 1's in that block is the value of the function computed at the argument. If M never halts after being starting on an argument, or it halts in a position other than the one just specified, the function computed has no value at that argument.

As I mentioned earlier, this simple formalization of the notion *computable function* is equivalent to recursive and λ-definable functions, and, so, is very powerful. Nevertheless, it is easy to see that there are total functions, from positive integers to positive integers, that are not Turing-computable. Because each machine can be regarded as a finite sequence of symbols of an enumerable "alphabet," where each such sequence meets a few specifiable conditions, it is possible to enumerate the set of Turing machines, effectively assigning each machine a unique index. Since each machine computes a total or partial function from positive integers to positive integers, this means we can enumerate all such Turing-computable functions—f_1, f_2, f_3, . . . [41] Next we define a function g such that $g(n) = 1$ iff $f_n(n)$ is undefined, while $g(n) = 1 + f_n(n)$ iff f is defined at argument n. If g is Turing-computable, it must have at least one index m. When we ask for its value at m, we see that if g is undefined at m, then $g(m) = 1$, which is impossible, and if $g(m) = k$ iff $g(m) = k+1$, which is impossible. Thus g is not Turing-computable. Assuming that all intuitively computable functions

[41] Boolos and Jeffrey (1974), pp. 17–18 and 45–47. Since distinct machines can compute the same function, some functions may occur more than once on the list. This doesn't affect the point.

are Turing-computable (or equivalently recursive or λ-definable), we conclude that g isn't computable.

This elementary result is closely related to the *halting problem for Turing machines*. For each machine M and number n, we ask whether M will eventually halt after being started on a tape (in standard initial configuration) representing n (i.e., a block of n squares with dots on an otherwise blank tape). Is there a decision procedure which, given any Turing machine and argument n, will always correctly tell us in a finite number of steps whether or not the machine will eventually halt? Although the answer to this question may not appear obvious, it is demonstrated in Turing (1936/37) not to be answerable by a Turing machine. So, if, as Church and others have conjectured, any intuitively computable function is computed by a Turing machine, then the halting problem for Turing machines is absolutely unsolvable.

In order to appreciate the result, one must grasp the significance of the question *Is the halting problem for Turing machines solvable by a Turing machine*? How, one might ask, can the question even be asked? Since the only Turing computable functions we have spoken of have mapped numbers onto numbers, one may wonder what it would mean for one Turing machine to take another Turing machine M plus an arbitrary positive integer as arguments, and assign one value—say the number 2—iff M eventually halts after being started on that input, and another value—say the number 1—iff M never halts. The answer should, by now, be familiar. We make sense of the idea by, in effect, Gödel-numbering Turing machines. Since each can be regarded as a well-formed word in a symbolic language, we can numerically code each M in a way that allows us to recover M's instructions from its code. Using this method, we can, in principle, construct a universal Turing machine U. Given a pair consisting of the numerical code of an arbitrary machine M plus its argument n, U reproduces M's operations on n.

So, is the halting problem solvable by a Turing machine? Let, M_1, M_2, M_3, . . . be an enumeration of all and only Turing machines in which each machine occurs at just one place in the list. (Assume we can recover the Gödel number of each machine, and hence its instructions, from its numerical index in the list.) Let h be a two-place halting function such that for any natural numbers m,n, h(m,n) = 2 iff the m^{th} Turing machine in our enumeration eventually halts after being started on n as input, and otherwise h(n) = 1. This function will be computed by a Turing machine, HALT, only if another Turing Machine, HALT*, constructible from it, halts on any input n iff the n^{th} Turing machine in our list fails to halt on its own index n. Of course, if there were such a machine as HALT*, it would have its own index n*. We would then have the contradictory result that HALT* eventually halts after being started on n* iff it fails to halt. Thus, there must be no Turing machine HALT that computes the halting function for Turing machines.

The proof is easy.[42] Supposing the existence of HALT, we construct HALT* consisting of HALT plus two other Turing machines, COPY and STOP/SPIN. Given an input of a single block of n squares with dots on an otherwise blank tape, COPY outputs a tape in which the initial block is followed by a blank square followed by a copy of the first block. This is inputted to HALT, which, by hypothesis, produces a tape with a single block of two squares with dots, *if the nth machine halts after being started on n,* or a tape with a single dotted square, *if the nth machine never halts on that input*. Given the former input, STOP/SPIN puts HALT* into an unending loop of moving right that never halts; given the latter input, the machine checks to see there is only one dot and halts in standard position. Hence HALT* fails to halt when started on its own index iff it does halt when started on its index. Since this is impossible, neither HALT* nor HALT exists. So, the halting problem for Turing machines is not solvable by a Turing machine.

5.3. Undecidability via the Halting Problem

Whereas Church's proof of the undecidability of first-order logic is built on Gödel's first incompleteness theorem, Turing's proof shows that if there were a decision procedure for first-order logical consequence, then there would be a solution to the halting problem. Since there is no solution to the latter, there is no decision procedure for the former. The trick to reducing the halting problem to the decision problem for first-order logic is to show how, given any Turing machine M and input n, one can effectively construct a sentence of the first-order predicate calculus that will be logically true iff M eventually halts after being started on n.

The sentence $\ulcorner(D \rightarrow H)\urcorner$ we need is one the antecedent D of which (at some interpretation I) both describes the tape on which M starts working and encodes all M's instructions, while H is a sentence true (at I) iff M eventually halts. So, if $\ulcorner(D \rightarrow H)\urcorner$ is logically true (i.e., if H is a logical consequence of D), then all models of D (including I) are models of H, which means that H is true (at I), guaranteeing that M will eventually halt. Conversely, suppose M will eventually halt after being started on a tape at t_0 correctly described by D (as interpreted at I). Then, since M's subsequent action is determined by the input plus the instructions encoded by D, for each time t_i there will be a logical consequence of D which (interpreted at I) correctly describes the tape at t_i. Since H is one of these consequences, H is a logical consequence of D, if M eventually halts. In short, deciding whether $\ulcorner(D \rightarrow H)\urcorner$ is logically true will decide whether M eventually halts.

Here are a few of the details. Think of the squares on the tape as numbered, with positively numbered squares to the right of square zero and

[42] Ibid., pp. 49–50.

negatively numbered squares to the left. M starts at time zero on square zero. The domain of interpretation I includes all positive and negative integers plus zero. The first-order language for our sentence ⌜(D → H)⌝ includes the name '0', which (in I) designates zero; finitely many two-place predicate letters Q_i (for each of M's finitely many internal states) that are true (at I) of pairs of numbers x,y such that at time x M is in state i scanning square number y; two two-place predicate letters S_0 and S_1 that are true (at I) of pairs numbers x,y iff at time x square number y is blank (if the symbol is S_0) or has a dot (if the symbol is S_1); two more two-place predicate letters *R* and *L, the former being true (at I) of pairs of numbers x and y iff at time x M is scanning the square immediately to the right of y (the number of which is one more than that of y), the latter being true of pairs of numbers x and y iff at time x M is scanning the square immediately to the left of y (the number of which is one less than that of y)*; the two-place predicate '<' that is true (at I) of x,y iff x is less than y; and, finally, a symbol '*' we will use for the successor function.

Instructions look like (i)–(vi).

(i) $\forall t\forall x\forall y\,[(tQ_ix \,\&\, tS_0x) \rightarrow (t^*S_1x \,\&\, t^*Q_kx \,\&\, (x{\neq}y \rightarrow ((tS_0y \rightarrow t^*S_0y) \,\&\, (tS_1y \rightarrow t^*S_1y))]$

When M is in state i scanning a blank square, it prints a dot in the square and goes into state k, leaving all other squares untouched.

(ii) $\forall t\forall x\forall y\,[(tQ_ix \,\&\, tS_1x) \rightarrow (t^*S_0x \,\&\, t^*Q_kx \,\&\, (x{\neq}y \rightarrow ((tS_0y \rightarrow t^*S_0y) \,\&\, (tS_1y \rightarrow t^*S_1y))]$

When M is in state i scanning a square with a dot, it erases the dot and goes to state k, leaving all other squares untouched.

(iii) $\forall t\forall x\forall y\,[(tQ_ix \,\&\, tS_1x) \rightarrow (t^*Rx \,\&\, t^*Q_kx \,\&\, ((tS_0y \rightarrow t^*S_0y) \,\&\, (tS_1y \rightarrow t^*S_1y))]$

When M is in state i scanning a square with a dot, it moves one square to the right and goes into state k, leaving all squares on the tape untouched.

(iv) $\forall t\forall x\forall y\,[(tQ_ix \,\&\, tS_1x) \rightarrow (t^*Lx \,\&\, t^*Q_kx \,\&\, ((tS_0y \rightarrow t^*S_0y) \,\&\, (tS_1y \rightarrow t^*S_1y))]$

When M is in state i scanning a square with a dot, it moves one square to the left and goes into state k, leaving all squares on the tape untouched.

(v) $\forall t\forall x\forall y\,[(tQ_ix \,\&\, tS_0x) \rightarrow (t^*Rx \,\&\, t^*Q_kx \,\&\, ((tS_0y \rightarrow t^*S_0y) \,\&\, (tS_1y \rightarrow t^*S_1y))]$

When M is in state i scanning a blank square, it moves one square to the right and goes into state k, leaving all squares on the tape untouched.

(vi) $\forall t\forall x\forall y\,[\,(tQ_ix \,\&\, tS_0x) \rightarrow (t^*Lx \,\&\, t^*Q_kx \,\&\, ((tS_0y \rightarrow t^*S_0y) \,\&\, (tS_1y \rightarrow t^*S_1y))]$

When M is in state i scanning a blank square, it moves one square to the left and goes into state k, leaving all squares on the tape untouched.

Sentence D, in ⌜(D → H)⌝, will contain a statement of this sort for each of the finitely many instructions of machine M. D will also contain a description of the initial configuration IC in which M starts. When that configuration consists of a single block of squares with dots on an otherwise blank tape, with M in state 1 scanning square zero, IC is the sentence that (interpreted at I) describes M's starting position.

IC. $0Q_10 \;\&\; 0S_10 \;\&\; \ldots \;\&\; 0S_10^{**\,(\text{n-1 times})} \;\&\; \forall y\,[(y\neq 0 \;\&\; y\neq 0^* \;\&\; \ldots \;\&\; y\neq 0^{**\,(\text{n-1 times})}) \rightarrow 0S_0y]$

M starts scanning the leftmost dotted square on a tape that consists of one block of n consecutive dotted squares, with all the other squares blank.

Since *entailment* of H by D requires *proving* some inequalities of the sort *0**(n times) ≠ 0** (n-1 times),* D must also include enough arithmetic, e.g., (vii)–(ix), to do so.

(vii) $\forall x\forall y\forall z\,((x < y \;\&\; y < z) \rightarrow x < z)$
(viii) $\forall x\forall y\,(x < y \rightarrow x \neq y)$
(ix) $\forall x\forall y\,(x^* = y \rightarrow x < y)$

Finally, we need a first-order sentence H which, when interpreted at I, is true iff M eventually halts. M will halt iff there comes a time t at which M is in a state represented by Q_j scanning a blank or a dotted square, represented by S_0 or S_1 respectively, and there is *no* instruction for M telling it what to do in that situation. So we inspect M's instructions, looking for one with a Q_j representing one of its internal states and an S_i representing one of the two symbols it can read, such that there is no instruction starting with the pair Q_j,S_i telling M what to do in that situation. If M ever halts, it will halt because it will have reached a point in its computation corresponding to such a pair. Of course, if there are any such pairs, there will only be finitely many, in which case we let H be the disjunction of corresponding sentences of the following form:

H. $\exists t\exists x\,(tQ_jx \;\&\; tS_ix)$

If, for a given machine, there are no such pairs, we know it will never halt no matter what input it is given, and so we let H be '0≠0'.

This completes the construction of the crucial sentence $\ulcorner(D \rightarrow H)\urcorner$ that is logically true iff H is a logical consequence of D iff machine M will eventually halt on an arbitrary input n.[43] If there were an effective procedure for determining whether H was a logical consequence of D, then there would be an effective procedure for determining whether or not M will eventually halt on input n. Since the construction of $\ulcorner(D \rightarrow H)\urcorner$ can be effectively carried out for arbitrary Turing machines, a decision procedure for first-order logical truth would ensure a decision procedure for solving the halting problem for Turing machines generally. On the assumption that the halting problem is unsolvable, it follows that first-order logic is undecidable—i.e., there is no decision procedure for first-order logical truth, or logical consequence. This is the Turing version of Church's

[43] The verification that $\ulcorner(D \rightarrow H)\urcorner$ is logically true iff M eventually halts on M is carried out in detail in Boolos and Jeffrey (1974), chapter 10, pp. 119–22. The first eight chapters of that book provide clear and informative background.

Theorem. From Gödel's completeness theorem for first-order logic, we know that there is an effective positive test for logical truth, and consequence. We now see that there is no effective negative test.

6. LEGACY

The decade between Gödel's dissertation in 1929 and the start of World War II was the most remarkable period ever in the history of symbolic logic, the foundations of mathematics, and the development of the mathematical theory of computation. During this period, much of what started with philosophically minded mathematicians and mathematically minded philosophers—Frege, Russell, Cantor, Zermelo, Hilbert, Gödel, Tarski, Church, and Turing, among others—came to maturity. The ambitious project in the philosophy of mathematics launched by Frege and Russell led not only to new departures in philosophical logic and the philosophy of language, but also to new deductive disciplines—of set theory, model theory, proof theory, recursive (computable) function theory, and the like. These, in turn, laid the foundations of the digital age that has, by now, transformed so much.

Two features of Turing machines were particularly important in this respect. First, they were digital, operating on zeros and ones. Thus they were perfectly suited to model the two positions, open and closed, of an electric circuit, thereby immediately suggesting the possibility of electronic computing machines. Second, their instructions, which determine their operation at every moment, can be encoded in formal languages, including the first-order predicate calculus. Thus they were capable of extracting logical consequences of linguistically encoded information of enormous variety.

Finally, the philosophers discussed here also helped set the stage for new conceptions of language, mind, and information advanced by their philosophical successors, who have made, and continue to make, foundational contributions to emerging sciences in these domains. In a discipline so accustomed to overreach and disappointment, it is worth emphasizing that stunning success is possible.

CHAPTER 9

Tarski's Definition of Truth and Carnap's Embrace of "Semantics"

1. BACKGROUND

In the mid-1930s, Alfred Tarski published two articles that soon became classics. In 1935 he published "The Concept of Truth in Formalized Languages," in which he defines truth for formal languages of logic and mathematics.[1] In 1936, he published "On the Concept of Logical Consequence," in which he uses the definition of truth to provide the basis for the now standard "semantic," or model-theoretic, definition of logical consequence

[1] Tarski (1935) is the German translation (plus an added postscript) of the original Polish version published in 1933. The English translation appears in Tarski (1983).

(and related notions).[2] Tarski's interest in truth arose from an interest in the expressive power of mathematical languages and theories, including the *definability* of important metatheoretical notions in them. To say that a set s is definable in a language L is to say that there is some formula F(v) of L (with free variable v) that is *true of* all and only the members of s. Tarski (1931 [1983]), "On Definable Sets of Real Numbers," investigates a language L sufficient for formulating a theory of the arithmetic of the real numbers. There, he gives a recursive characterization of *the set of real numbers definable in L*, using ideas that received fuller expression in Tarski (1935). In addition, his proof that the set of Gödel numbers of true sentences in the language LA of arithmetic is not definable in LA (explained in section 2 of chapter 8) is also given in Tarski (1935).

Tarski's investigations of definability and truth began in seminars he gave at the University of Warsaw between 1927 and 1929.[3] Since there were then no definitions of these concepts in terms of the concepts of logic and set theory used in metamathematical investigations, and also no axiomatic theory of *truth* and *definability* taken as primitives, mathematicians regarded them with some suspicion. The designation of these notions as "semantic" didn't help, in part because of their role in paradoxes like the Liar, and in part because it wasn't obvious how the concept *true sentence* could be treated with the same rigor and formality as, for example, the recursive concept, *proof* (in a given system) and the recursively enumerable concept, *provable* (in a given system). Nevertheless, Tarski rightly regarded both *truth* and *definability* to be essential to metamathematics, which led him to try to make them respectable.

He displays this dual attitude concerning semantical notions—regarding them as metamathematically needed while themselves requiring mathematical reconstruction and validation—in Tarski (1931 [1983]) and Tarski (1935). In Tarski (1931 [1983]), he notes the need to overcome mathematicians' skeptical attitude toward *definability*.

> The distrust of mathematicians towards the notion in question is reinforced by the current opinion that this notion is outside the proper limits of mathematics altogether. The problems of making its meaning more precise, of removing the confusions and misunderstandings connected with it, and of establishing its fundamental properties belong to another branch of science—metamathematics.[4]

In Tarski (1935), he emphasizes the metamathematical importance of semantic concepts by citing the dependence on them of such central results as Gödel's completeness theorem for the first-order predicate calculus

[2] Tarski (1936) is the German version of a lecture delivered in 1935. The English translation appears in Tarski (1983).
[3] See Vaught (1974, 1986).
[4] Tarski (1931 [1983]), p. 110.

and different versions of the Lowenheim-Skolem theorem.[5] About these results, he says:

> [I]t is evident that all these results only receive a clear content and can only then be exactly proved, if a concrete and precisely formulated definition of [true] sentence is accepted as a basis for that investigation.[6]

The task of Tarski (1935) is to provide a formal definition of *truth* for the languages of logic and mathematics, which will in turn make it possible to formalize and vindicate other needed semantic concepts, including *definability* and *logical consequence*.

2. TRUTH, PARADOX, AND INCONSISTENCY

Tarski's first task was to insulate the formal truth predicate he wished to define from doubts about the coherence of our informal concept of truth stemming from the Liar paradox. For that reason, he sought to identify the features of the ordinary truth predicate responsible for the paradox, and to exclude them from his definition. To understand his diagnosis and its consequences for the definition he sought to provide, one must start with some basic observations about truth predicates in natural languages.

The English predicate *is true* can be applied not only to sentences, but also to statements, propositions, and uses of sentences (if these are distinct). It correctly applies to a sentence S only if S is used to make a statement, or express a proposition, that is true. Although *is a true sentence* is an English predicate, its application is universal. It applies to any sentence of any language that is used to make a true statement (or express a true proposition), and to only such sentences. The related predicate *is true* is capable of applying to any statement/proposition one might make or express, and to any sentence used to make or express it. Since *is true* is itself used to make or express statements/propositions, it is applicable to the statements/propositions it is used to make or express, and to the sentences used in doing so. This leads to the Liar paradox.

There are many versions of the paradox involving sentences or propositions that seem to say of themselves that they aren't true. Tarski was concerned with sentential versions built around Liar-sentences like (1).[7]

[5] See Tarski (1935 [1983]), pp. 240–41.
[6] Ibid., p. 241.
[7] Propositional versions of the paradox differ significantly from sentential versions, but they were not much studied at the time—due to the lack of any viable conception of non-linguistic propositions and to the then widespread tendency to slight important differences between direct and indirect discourse reports.

1. Sentence (1) is not true.

We stipulate that the expression *sentence (1)* is here used as an abbreviation of the singular term *the sentence that is the first numbered example in section 1.1 of chapter 9 of volume 2 of The Analytic Tradition in Philosophy.* So understood, (1) is a meaningful sentence of English, as is shown by the fact that someone not familiar with this book would easily understand it. In addition, *what (1) says*, the proposition it expresses, would have been true if the first numbered example in this section had been *There are no even prime numbers.* Since sentence (1) is used to say something that would have been true had certain circumstances obtained, it must be meaningful.

Despite this, it is paradoxical because a contradiction can be derived from apparently incontrovertible assumptions about it.

VERSION 1 OF THE LIAR

P1. 'Sentence (1) is not true' is a true iff sentence (1) is not true.
P2. Sentence (1) = 'Sentence (1) is not true'.
C1. Sentence (1) is true iff sentence (1) is not true.
C2. Sentence (1) is true and sentence (1) is not true.

C1 is derived by substituting the expression *Sentence (1)* for the quotename '*Sentence (1) is not true*' in P1 on the basis of P2, which, if true, ensures that these expressions are coreferential. Given that the linguistic context *x is true iff sentence (1) is not true* is extensional, we derive C1 from P1 and P2. Given that C2 is a tautological consequence of C1 in the classical propositional calculus, we derive C2.

Having validly derived a contradiction from P1 and P2, we must reject one or the other. But we can hardly reject P2, which can be established merely by inspecting (1) above. Rejecting P1 is also difficult. Since P1 is an instance of *Schema True*, its correctness seems to be guaranteed by our notion of truth.

Schema True: 'X' is a true sentence of English iff P (where 'P' is replaced either by the sentence S replacing 'x' or by a sentence synonymous with S).

How could any instance of this schema be false? A claim ⌜P iff Q⌝ can be false only if P is true and Q is false, or Q is false and P is true.[8] But when P is ⌜'A' is true⌝ and Q is A, these combinations seem impossible. Surely the claim *that A is true* can't be true when A is false, nor can the claim *that A is true* be false when A is true. But if no instance of *Schema True* can be denied, then P1 can't be denied. This is the paradox. Although we have derived a contradiction, which must be rejected, we can't see how to do so because both our premises and our logic seem unassailable.

[8] The symbols 'P', 'Q', and 'A' are used as metalinguistic variables over sentences in this paragraph.

The problem arises from the natural thought that *Schema True* incorporates a linguistic rule essential to understanding the truth predicate. Suppose one were asked to explain the meaning of *is true* to someone who knew a lot of English but wasn't yet acquainted with the word *true* or any synonym. Suppose one's goal was to explain its application to sentences of English. One could hardly do better than to say something like this:

> "The sentence *snow is white* is true iff snow is white, the sentence *the sun shines nearly every day in Seattle* is true iff the sun shines nearly every day in Seattle, the sentence *there is a duplicate of the earth somewhere in the Milky Way* is true iff there is a duplicate of the earth somewhere in the Milky Way, and so on for every meaningful declarative sentence of English."

Here, it seems, one explains what the truth predicate means by conveying the correctness of all instances of *Schema True*. If so, how can one who understands the predicate justifiably reject any such instance?

Although these considerations might make one skeptical about the concept expressed by *is true,* it is worth noting that nothing in our statement of the Liar is, by itself, an attack on the legitimacy or coherence of that concept. So far, we simply have a deep and perplexing puzzle—to which many different solutions have been proposed. Nevertheless, one can understand how the considerations just rehearsed *might* lead one to suspect that the concept of truth, as we ordinarily understand it, is defective. One line of reasoning leading to this conclusion is this: The ordinary concept of truth *requires* all instances of *Schema True* to be true, which the Liar paradox shows to lead to contradiction. Therefore, the concept expressed by *is a true sentence of English* is incoherent; its presence is a defect of the language that needs to be rectified.

Tarski suggests something close to this line of reasoning in Tarski (1935).

> A characteristic feature of colloquial language (in contrast to various scientific languages) is its universality. It would not be in harmony with the spirit of this language if in some other language a word occurred which could not be translated into it; it could be claimed that 'if we can speak meaningfully about anything at all, we can also speak about it in colloquial language'. If we are to maintain this universality of everyday language in connexion with semantical investigations, we must, to be consistent, admit into the language, in addition to its sentences and other expressions, also the names of these sentences and expressions, and sentences containing these names, as well as such semantic expressions as 'true sentence', 'name', 'denote', etc. But it is presumably just this universality of everyday language which is the primary source of all semantical antinomies, like the antinomies of the liar or of heterological words. *These antinomies seem to provide a proof that everyday language which is universal in the above sense, and for which the normal laws of logic hold, must be inconsistent.* This applies especially to the formulation of the liar which I have given. . . . If we analyze this antinomy . . . *we reach the conviction that no consistent language*

can exist for which the usual laws of logic hold and which at the same time satisfies the following conditions: (I) for any sentence which occurs in the language a definite name of this sentence also belongs to the language; (II) every expression formed from (2) [*x is a true sentence iff p*] by replacing the symbol 'p' by any sentence of the language and the symbol 'x' by a name of this sentence is to be regarded as a true sentence of the language; (III) in the language in question an empirically established premise having the same meaning as (α) [*'c is not a true sentence' is identical with c*] can be formulated and accepted as a true sentence.[9]

Tarski's argument has the following structure:

A. Any language that satisfies certain conditions is inconsistent.
B. English and other natural languages satisfy these conditions.
C. Therefore, English and other natural languages are inconsistent.

His conditions appear to be:

0. The usual laws of logic hold in the language.
.5 The language contains a truth predicate applying to its own sentences.
I. The language contains names of all its sentences.
II. All instances of *Schema True* are true sentences of the language.
III. An empirical premise analogous to P2 above is a true sentence of the language.

It should be noticed at the outset that the above formulation of conditions (II) and (III) is somewhat more definite than Tarski's more equivocal wording. Instead of flatly saying that each instance of Schema True is a true sentence of the language, he says that each instance "is *to be regarded as* a true sentence of the language." Instead of saying that an empirical premise analogous to P2 is a true sentence of the language, he says that it "*can be formulated and accepted as* a true sentence of the language." Although this evasive wording doesn't convey any clear alternative to the straightforward formulations (II) and (III), there is, as we shall see, reason to think that Tarski may have felt uncomfortable with the unequivocal formulations, without knowing what to replace them with. I will return to this idea after reconstructing the argument based on the conditions stated above.

Condition (0) may be assumed to guarantee the existence of all standard logical vocabulary in the language, including classical negation. This plus condition (.5) guarantees that the language contains what we may informally call an *untruth predicate* that applies to one of its sentences iff it is not true. Condition (I) must be interpreted liberally. If the only way of talking about sentences was by using their quote-names, then paradoxical sentences like sentence (1) would not be available. That isn't what Tarski

[9] Tarski (1935 [1983]), pp. 164–65, my emphasis.

had in mind. Instead, he assumed that sentences of the language are freely nameable, describable, and/or quantified-over using arbitrary names, complex singular terms (e.g., Fregean definite descriptions), or ordinary quantifiers (guaranteed by condition (0)). If these conditions are met, the language will contain paradoxical sentences that can play the role of sentence 1 in version 1 of the Liar.

This gives us assumption 1 of version 2 of the Liar, which is reconstructable from Tarski's remarks.

A1. 'Sentence (1) is not true' is a sentence of English.

The next two assumptions, given by conditions (II) and (III), are:

A2. All instances of *Schema True* are true sentences of English.
A3. The sentence *Sentence (1) = 'Sentence (1) is not true'* is a true sentence of English.

Note the relationship between A1–A3 and assumptions P1 and P2 of version 1 of the Liar. Whereas P1 is an instance of *Schema True*, A2 is the metalinguistic claim that all instances of *Schema True* are true. Since A1 ensures that P1 is such an instance, A1 and A2 yield the result that sentence P1 is true. A3 is the claim that sentence P2 is also true. Thus, version 2 of the Liar asserts that the premises of version 1 are true sentences of English. Since they were used to derive a contradiction, we may expect version 2 to reach the conclusion that English contains a true contradiction.

To get this result, we need to say more about what is meant by Tarski's condition that the usual laws of logic hold in English. Presumably, it means that all classically valid inferences preserve truth in English.[10] On this interpretation, one who asserts that the usual laws of logic hold in English assumes (i) that certain English expressions correspond to operators in classical logic, (ii) that certain sequences of English sentences can be recognized as instances of logically valid patterns of inference, and (iii) that if a sequence $S_1 \ldots S_n$ of English sentences corresponds to the premises of a logically valid argument $P_1 \ldots P_n / C$, and S_c is an English sentence that corresponds to C, then S_c will be true sentence of English if $S_1 \ldots S_n$ are also true. Incorporating this understanding of condition (0) into a further Tarskian assumption A4 will allow us to derive the conclusion that the contradictory C2 (of version 1 of the Liar) is a *true* sentence of English.

A4. The usual laws of logic hold in English—i.e., all classically valid patterns of inference are truth-preserving in English.

To complete the reconstruction of Tarski's remarks, we need an interpretation of the puzzling notion of an *inconsistent language*. We don't normally

[10] This way of understanding the condition was originally suggested to me by my former colleague, Nathan Salmon, in my Princeton seminar on truth on April 2, 1981.

think of languages as the kinds of things that can be consistent or inconsistent. Certainly the existence of inconsistent sentences in a language isn't enough to make it inconsistent in any pejorative sense. After all, any language with negation contains inconsistent sentences, and any language with both negation and conjunction contains contradictory sentences. Far from being a defect, which Tarski takes inconsistency to be, having the resources to construct inconsistent sentences is a virtual prerequisite for a language to be rich enough to be interesting. Since A1–A4 will allow us to derive the conclusion that English contains true sentences that are either themselves inconsistent or inconsistent with other true sentences, we may take Def to define what it is for a language to be inconsistent.[11]

Def. A language is inconsistent iff some sentence and its negation are true sentences of the language.

This gives us version 2 of the *Liar*.

VERSION 2 OF THE LIAR

A1. 'Sentence (1) is not true' is a sentence of English.
A2. All instances of *Schema T* are true sentences of English.
A3. The sentence *Sentence (1) = 'Sentence (1) is not true'* is a true sentence of English.
A4. The usual laws of logic hold in English—i.e., all classically valid patterns of inference are truth-preserving in English.
C1. ''Sentence (1) is not true' is true iff sentence (1) is not true' is a true sentence of English. (From A1 and A2)
C2. 'Sentence (1) is true iff sentence (1) is not true' is a true sentence of English. (From C1, A3, and A4's guarantee that substitutivity of identity is truth-preserving)
C3. 'Sentence (1) is true and sentence (1) is not true' is a true sentence of English. (From C2 and A4's guarantee that tautological consequence is truth-preserving)
C4. 'Sentence (1) is true' and 'Sentence (1) is not true' are true sentences of English. (From C3 and A4's guarantee that simplification of conjunction is truth-preserving)
A5. 'Sentence 1 is not true' is a negation of 'Sentence 1 is true'.
Def. A language is inconsistent iff some sentence and its negation are true sentences of the language.
C5. English is an inconsistent language.

This metalinguistic version of the Liar extractable from Tarski's remarks parallels the non-metalinguistic version 1 of the Liar. Does it call for a

[11] Salmon also made this suggestion in the seminar mentioned in note 10. Other proposals for interpreting the notion of an inconsistent language are discussed in Soames (1999), pp. 62–64, fn. 53.

similar response? In version 1, we derive a contradiction. Assuming that we can't rationally accept a contradiction, we must reject at least one premise or rule of inference used in the derivation. Although in version 2 we don't derive a contradiction, we do derive the conclusion that a contradiction is true. Isn't this conclusion as undesirable as the conclusion of version 1?[12] If so, then here too we must reject a premise or a rule of inference. But if we do that, we can't claim to have shown that English is an inconsistent language, or that the ordinary truth predicate it contains is incoherent. Instead, we must see the paradox not as a source of truth-skepticism, but as a demonstration that at least one initially plausible assumption about the ordinary concept of truth is incorrect.

This was not Tarski's response. Whereas he clearly didn't accept all premises, rules of inference, and conclusions of version 1 of the Liar, he was apparently willing to do so for version 2. In Tarski (1944) he says this about version 1.

> In my judgment, it would be quite wrong and dangerous from the standpoint of scientific progress to depreciate the importance of this and other antinomies. . . . It is a fact that we are here in the presence of an absurdity, that we have been compelled to assert a false sentence (since (3) [*'S' is true iff 'S' is not true*], as an equivalence between two contradictory sentences, is necessarily false). If we take our work seriously, we cannot be reconciled with this fact. We must discover its cause, that is to say, we must analyze premises upon which the antinomy is based; we must then reject at least one of these premises, and we must investigate the consequences which this has for the whole domain of our work.[13]

Instead of rejecting paradox-creating assumptions about English (and other natural languages) used in version 2 of the Liar, Tarski rejected the languages themselves as inadequate for the construction of serious theories of truth, and proposed they be replaced for these purposes by formalized languages for which restricted truth predicates can be defined that make the construction of Liar sentences impossible.

This position is questionable. It is hard to accept the claim that a contradictory sentence of English is true, or that English contains a sentence and its negation that are jointly true. One reason these claims are difficult to accept may be that our understanding of *true* and *negation* precludes the possibility of any true sentence S* counting as the negation of another true sentence S. The centrality of disquotation to our understanding of the ordinary sentential truth predicate is another reason. Disquotation, which allows one to pass from ⌜'S' is true⌝ to S, and conversely, is an extremely useful feature of the truth predicate, allowing us, for example, to express generalizations like *Every meaningful sentence*

[12] Given A4, we can, of course, also derive that every sentence of English is true.
[13] Tarski (1944 [1952]), p. 20.

of the form 'If S, then S' is true, which commit us to each member of a class of statements of the form *'If S, then S'*, which is too large to list. But one who accepts the argument in version 2 of the Liar, and says, in English, that a certain contradictory sentence of English is true, can avoid contradicting himself only by *denying* the disquotational inference (from a metalinguistic truth claim to a non-metalinguistic claim). This denial is made all the more unpalatable when one considers the rationale for the crucial assumption A2 of version 2—the assumption that all instances of *Schema True* are true. Surely, the pretheoretic conviction that instances of this schema are *true* is based on the conviction that they are correctly *assertable*. Thus, once the proponent of the argument in version 2 *rejects* the assertability of some instances of *Schema True*, as Tarski apparently does, he forfeits his strongest, and perhaps only, pretheoretic warrant for accepting A2.

Because of this, it is reasonable to suppose that there may be other, more plausible diagnoses of what the Liar paradox shows about natural language—diagnoses that involve rejecting one or more of the premises and rules of inferences used in the paradoxical arguments—and replacing them with otherwise acceptable substitutes that don't lead to paradox. But this is not our concern. What is important for us is both that Tarski recognized the need to insulate the truth predicates he required for his metatheoretical investigations from paradox, and that he succeeded in doing so. He did this by using a metalanguage M to specify a formal object language L, and then defining, in M, a restricted truth predicate T applying to sentences of L. Since L doesn't itself contain a truth predicate, no Liar-paradoxical sentences are constructible in it. Since any sentence S of M containing the predicate T is not a sentence of L, the truth predicate contained in S doesn't apply to S itself. Thus S can't be seen as asserting or denying its own truth. If, one wants a truth predicate for M, the process can be repeated in a higher metalanguage M+.

Tarski's reasons for wanting to avoid the Liar paradox, and his success in introducing metamathematically acceptable truth predicates that do so, were more important than any diagnosis of the status of ordinary truth predicates lurking in his brief remarks on the subject. One who wishes to use truth as a precise metamathematical instrument needs to put the persistent confusion engendered by the Liar paradox behind him, which Tarski did. That said, his idea that languages—which are not formal theories and don't themselves assert anything—can be inconsistent has been more trouble than it is worth. Although he did seem to think that ordinary truth predicates are defective, he neither successfully articulated what the defect was supposed to be, nor, apparently, put much effort into doing so. Wishing to avoid the ordinary notion of truth, it was natural that he should feel uneasy about using it himself in giving his argument that languages satisfying certain conditions are inconsistent. Hence his evasive language about instances of the truth schema being "regarded as true" and the empirical premise being "accepted as true."

The argumentative strategy behind his use of this language was, I think, to put the onus on anyone who finds the ordinary notion of truth acceptable, and not in need of replacement for metamathematical work. *Very well*, Tarski implicitly whispers, *if you don't see the need for a formal definition of truth of the sort I am proposing, tell me what you reject. Is it the ability to construct and identify Liar sentences? Is it the laws of logic? Is it some, or all, instances of the truth schema? You need to tell me, for if you don't reject any of them, then you—not the language you use—will be committed to the idea that your truth predicate applies to contradictions (and indeed to every sentence) and so is useless for scientific purposes. In the meantime, I have serious work to do*. The virtue of this message is that it could have been effective without causing further confusion and raising unresolved issues about whether it is our ordinary truth predicates that are defective, or whether it is we ourselves who are, because we don't yet have a non-defective theory of them. Would that Tarski have left it at that.

3. TARSKI'S CRITERIA OF CORRECTNESS FOR DEFINING TRUTH

3.1. Material Adequacy and the Coextensiveness of Truth and Tarski-Truth over L

The languages Tarski was concerned with were of the first and higher-order predicate calculus. All expressions of each object language L for which he defines truth are assumed already to be understood by working logicians or mathematicians. This assumption is *not* gratuitous. His definition of truth does *not* provide an interpretation of sentences of L. Quite the opposite; the fact that those sentences are already used to make claims about a given domain of objects provides Tarski with the concept he wants the predicate introduced by his definition to express. As we will see, this is implicit in his criteria for determining when a proposed definition, in M, of truth for L is *correct*.

In order to make sure that the Liar paradox doesn't arise, Tarski stipulates that L doesn't itself contain predicates expressing semantic concepts (applying to arbitrary expressions of L); nor does it have the means of explicitly referring to, or quantifying over, arbitrary expressions or sets of expressions of L. It also doesn't contain indexical or context-sensitive expressions. The definition of the one-place predicate *true-in-L* is constructed in a richer metalanguage M that includes either the sentences of L themselves or translations of them. M also includes the resources needed to refer to, and quantify over, expressions (including sentences) and sets of expressions of L plus arbitrary sets of n-tuples of objects about which its sentences are used to make claims. When M and L are related in this way, Tarski shows how to construct an explicit definition in M of a predicate that applies to all and only the true sentences of L.

Before giving his definition, he lays down criteria for success. The most important criterion is that the definition be *materially adequate*. A definition in M of a predicate 'T_L' of sentences of L satisfies this criterion iff for

every sentence S of L, the definition entails an instance of *Schema T*, which is a sentence of M that results from replacing 'X' in *Schema T* with a name, N_S, of S and replacing 'P' with S itself (if S is also a sentence of M) or with a sentence, P_S, of M that is a *paraphrase* or *translation* of S.[14]

Schema T: X is T_L iff P

The role of *material adequacy* is to guarantee that the defined predicate 'T_L' is coextensive with 'is a true sentence of L', and so applies to all and only true sentences of L.

To illustrate this guarantee, I introduce *Schema TM*, instances of which are gotten by replacing 'X' with a transparent name of a sentence of L and 'P' with any sentence of M.

Schema TM: If X means in L that P, then X is true (in L) iff P

This schema connects our ordinary notions of truth and meaning. Typically, one who understands it is warranted in believing *that all its instances are true and assertable*. Let S be a sentence of L and ⌜N_S is T_L iff P_S⌝ be an instance of *Schema T* in which N_S names S. Since P_S means the same as S, the corresponding instance of TM ⌜If N_S means in L that P_S, then N_S is $true_L$ iff P_S⌝ has a true antecedent.[15] This gives us ⌜N_S is $true_L$ iff P_S⌝, which along with ⌜N_S is T_L iff P_S ⌝ allows us to derive ⌜N_S is T_L iff N_S is $true_L$⌝, *for every sentence of L*. Hence, we establish that if the definition of 'T_L' is materially adequate, then 'true' and 'T_L' are coextensive over L.

3.2. The Illusion that Truth and Tarski-Truth Are More Than Coextensive

For Tarski's metamathematical purposes, it is enough that 'T_L' be demonstrably coextensive with 'true' over L. But it is tempting to suppose, as he, Carnap, and others were tempted, that 'T_L' and '$true_L$' are also conceptually

[14] To say that the definition *entails* an instance *I* of *Schema T* is to say that *I* is a *logical consequence* of the definition (which is a universally quantified biconditional in M) plus (i) the statement in M of the syntax of L and (ii) some elementary set theory required for the definition. (If one likes one can include (i) and (ii) in the truth definition itself.) Tarski also requires the *names* of sentences of L that occur in instances of *Schema T* be transparent in the sense that the expression named can be identified simply by understanding the name. Quote-names are often assumed to be transparent, as are *structural descriptive names* such as *the expression that consists of the letter 'S' followed by the letters 'n', 'o', and 'w' in that order*. The idea is that anyone who understands both a sentence S of L and an instance of *Schema T* in which S is named should thereby know the instance to be true and be in a position to accept and assertively utter it.

[15] Here, and in much of what follows in the discussion of the relation of Tarski's truth predicate to our ordinary truth predicate, I include a subscript on '$true_L$' to remind the reader that we are considering our ordinary truth predicate restricted to the formal language L. Thus the subscripted predicate is to be understood along the same lines as the phrase 'is a true sentence of L'.

connected. The temptation can be illustrated by comparing the sentence pairs (2a,b) and (3a,b).

2a. 'John gave the book to Mary' is true_L iff John gave the book to Mary.
b. x knows *that 'John gave the book to Mary' is true_L iff John gave the book to Mary*.
3a. 'John gave the book to Mary' is T_L iff John gave the book to Mary.
b. x knows *that 'John gave the book to Mary' is T_L iff John gave the book to Mary*.

What does it take for one to know that (2a) is true and assertable, and for one to satisfy the predicate (2b)? It is tempting to suppose that merely understanding (2a) is enough. Now suppose that 'John gave the book to Mary' is a sentence of L and that 'T_L' is a Tarskian truth predicate for L defined in a metalanguage that contains L. What does it take for one to know that (3a) is true and assertable, and for one to satisfy the predicate (3b)? If (3a) is a consequence of the materially adequate definition of 'T_L', all it takes is for one to understand (3a) (which includes understanding the definition). Thus, just as understanding (2a) seems to warrant accepting it, and lead to one's satisfying (2b), so understanding (3a) warrants accepting it, and leads to one's satisfying (3b).

From this it might seem to follow that merely understanding (4a) warrants accepting it, and leads to one's satisfying the predicate (4b).

4a. 'John gave the book to Mary' is T_L iff 'John gave the book to Mary' is true_L.
b. x knows *that 'John gave the book to Mary' is T_L iff John gave the book to Mary' is true_L*.

Since this result—which doesn't depend on one's having any empirical information beyond that needed to understand 'true_L' and 'T_L'—can be repeated for every sentence of L, the coextensiveness of 'true_L' and 'T_L' over L may appear to be conceptually guaranteed. This is an illusion. But it is an illusion with a distinguished pedigree.

3.3. Tarski's Commitment to the Illusion

Tarski explained his definition of truth to a general philosophical audience in Tarski (1944), where he claims that his defined notion of *truth-in-L* is conceptually connected to our ordinary notion of truth, restricted to L. He says the definition "*does not aim to specify the meaning of a familiar word used to denote a novel notion; on the contrary it aims to catch hold of the actual meaning of an old notion*."[16] Since he took his truth predicate to capture what is essential to the ordinary predicate 'is a true sentence of L', he thought it could play all theoretical roles for which we might need

[16] Tarski (1944 [1952]), p. 13.

a notion of truth. Thus he says in section 13 that his notion of truth can be used to define semantic notions including *consequence, synonymy,* and *meaning*. He would not have said this had he not believed that his defined truth predicate for L comes very close to capturing the ordinary notion *being a true sentence of L*.[17]

His stance in Tarski (1969) is similar. First he explains what he calls *partial definitions of truth* (applying to individual sentences of a language); then he explains how a *general definition of truth* (for the language) is related to the partial definitions. Putting the Liar paradox aside, he uses English sentences as examples. He begins by discussing the meanings of sentences used to predicate truth or falsity of other sentences.

> Consider a sentence in English whose meaning does not raise any doubts, say the sentence 'snow is white'. For brevity we denote this sentence by 'S', so that 'S' becomes the name of the sentence. We ask ourselves the question: *What do we mean by saying that S is true or that it is false*? The answer to this question is simple: in the spirit of Aristotelian explanation, *by saying that S is true we mean simply that snow is white, and by saying that S is false we mean that snow is not white*.[18]

It is plausible that when one *says* that 'snow is white' is true *one means* or *asserts* that snow is white. But what Tarski seems to suggest by it, namely that the sentences *'snow is white' is true* and 'snow is white' *mean the same thing*, is not.[19] He would also have said that ⌜'snow is' is T_E⌝ means the same as, or at least is logically equivalent to, 'snow is white', when ⌜'snow is white' is T_E iff snow is white⌝ is a consequence of a materially adequate definition of a Tarskian truth predicate 'T_E' for a fragment E of English (constructed in a slightly extended fragment E+). Combining these results, we reach the conclusion that 'snow is white', *'snow is white' is true,* and ⌜'snow is white' is T_E⌝ are paraphrases of one another.[20] In short, ordinary truth, restricted to a language L, and Tarski truth for L are essentially the same thing.

Next, he describes (5a) and (5b) as *partial definitions* of truth and falsity.

> [5a] 'Snow is white' is true iff snow is white.
> [5b] 'Snow is white' is false iff snow is not white.

[17] Ibid., p. 28. Tarski attaches a footnote to the comment cited above about consequence, synonymy, and meaning. It directs the reader to Carnap (1942) for a definition of synonymy in terms of what Carnap took to be Tarski's notion of truth.

[18] Tarski (1969), p. 64, my emphasis.

[19] See Soames (2014), pp. 78, where it is argued that although the sentences mean different things, assertive utterances of either one typically assert both propositions.

[20] This combination of views is obviously problematic. Let 'T_S' be a Tarskian truth predicate for a fragment of Spanish containing 'la nieve es blanca', which means the same as the English sentence 'snow is white'. If 'snow is white' means both that *'la nieve es blanca' is T_S* and that '*snow is white' is true (in English),* then it would seem that the two metalinguistic sentences must mean the same thing. But surely they don't.

> Thus [5a] and [5b] provide satisfactory explanations of the meaning of the terms 'true' and 'false' when these terms are referred to the sentence 'snow is white'. *We can regard [5a] and [5b] as partial definitions of the terms 'true' and 'false'*, in fact as definitions of these terms with respect to a particular sentence. Notice that [5a], as well as [5b], have the form prescribed for definitions by the rules of logic, namely the form of *logical equivalence*. It consists of two parts, the left and the right side of the equivalence, combined by the connective 'iff'. The left side is the *definiendum*, the phrase whose meaning is explained by the definition; the right side is the *definiens*, the phrase that provides the explanation.[21]

In speaking of *meaning* and *definition,* Tarski was, I suspect, employing a conception of definition in which the *definiendum* D is taken to be an abbreviation of the *definiens* D*, which can be substituted for D in any formula in which it occurs. So, for any sentence provable using the definition, the corresponding sentence gotten by substituting D* for D is provable without the definition. Because of this, the definition doesn't increase the expressive power of any theory formulated in L. Hence, as Tarski insisted, his defined predicate 'T_L' can be eliminated from the metalanguage without loss of expressive power.

Next suppose that the sense in which he takes the *partial definition* (5a) to *give the meaning* of 'true' applied to 'snow is white' to be analogous to the sense in which instances of *Schema T* that follow from a materially adequate definition of 'T_E' give the meaning of Tarski's 'T_E' applied to individual sentences of the fragment E of English. For this to be so, it must be the case that *S is true iff S* is true, for any sentences S and S* of E in which S* is gotten from S by replacing occurrences of* "'snow is white' is true" *in S with occurrences of 'snow is white'* (and similarly for other sentences of E). But this result is questionable. Surely, there are English sentences in which substitution of 'snow is white' for "'snow is white' is true" fails to preserve truth. For example, (6a) may be true even if (6b) is not.

6a. In insisting that 'snow is white' is true, Bill was insisting that a three-word sentence of English was true.
6b. In insisting that snow is white, Bill was insisting that a three-word sentence of English was true.

These potential exceptions probably wouldn't have bothered Tarski. Since Tarski (1969) is an informal explanation of the strategy behind his definition for languages that don't contain hyperintensional, or even intensional, constructions, we can preserve our current understanding of *partial definition* by limiting its application to fragments E_L of English consisting of translations of sentences of formal languages L for which Tarski

[21] Tarski (1969), p. 64, my emphasis.

provides truth predicates. For every sentence S of E_L, there will be a "partial definition" ⌜'S' is true (in E_L) iff S⌝ (in a modest extension E+ of E) that Tarski would take to satisfy the italicized generalization above, while also taking it to be assertable by anyone who understood it. It is, I think, because he takes (5a,b) to be *partial definitions* in this sense that he suggests that the sentence "'snow is white' is true" *means that* snow is white, and the sentence " 'snow is white' is false" *means that* snow is not white.

This is the basis in Tarski (1969) for explaining the task of defining *true-in-L*. It consists of formulating a definition D the logical consequences of which include, for each sentence S of L, a *partial definition* of the defined predicate 'T_L'. That is what material adequacy comes to. As Tarski notes, in the imagined case of a language with only finitely many sentences, the problem of constructing such definition has a trivial solution.[22]

For example, let E be the fragment of English consisting of the following ten sentences.

1 is one of Bill's favorite numbers.
2 is one of Bill's favorite numbers.
.
.
.
9 is one of Bill's favorite numbers.
10 is one of Bill's favorite numbers.

Here is a trivial definition meeting Tarski's requirements.

7. For all sentences S of E, S is T_E (true in E) iff S = '1 is one of Bill's favorite numbers' and 1 is one of Bill's favorite numbers, or S = '2 is one of Bill's favorite numbers' and 2 is one of Bill's favorite numbers, or . . . or S = '9 is one of Bill's favorite numbers' and 9 is one of Bill's favorite numbers, or S = '10 is one of Bill's favorite numbers' and 10 is one of Bill's favorite numbers.

From the definition we derive (8).

8. '1 is one of Bill's favorite numbers' is T_E (true in E) iff '1 is one of Bill's favorite numbers' = '1 is one of Bill's favorite numbers' and 1 is one of Bill's favorite numbers, or '1 is one of Bill's favorite numbers' = '2 is one of Bill's favorite numbers' and 2 is one of Bill's favorite numbers, or . . . or '1 is one of Bill's favorite numbers' ='9 is one of Bill's favorite numbers' and 9 is one of Bill's favorite numbers, or '1 is one of Bill's favorite numbers' = '10 is one of Bill's favorite numbers' and 10 is one of Bill's favorite numbers.

Assume that we can derive each instance of the schema *'S' = 'S'* that results from replacing both occurrences of the letter 'S' with a sentence of E , and also derive each instance of *'S' ≠ 'S*'* that results from replacing the

[22] Ibid., p. 65.

occurrence of 'S' with a sentence of E and replacing the occurrence of the symbol 'S*' with a *different* sentence of E. Given this, we derive the *partial definition* (T1) from (8).

T1. '1 is one of Bill's favorite numbers' is T_E (true in E) iff 1 is one of Bill's favorite numbers.

Since *partial definitions* for the other nine sentences of E are similarly derivable, the definition is materially adequate, and so extends the several partial definitions to a materially adequate general definition. If each "partial definition" gives the meaning of the application of 'T_E' to a sentence of E *and that meaning is the same as the application of our ordinary predicate 'is a true sentence of E'*, then Tarski's defined predicate matches the meaning of the ordinary truth predicate over the language as a whole. That is the logic of the explanation presented in Tarski (1969).

The problem faced in Tarski (1935) was to find a way to reproduce this result for languages with infinitely many sentences. How, given such a language, can one extend all individual *partial definitions* of truth to a single, materially adequate general definition of truth? Tarski placed two further requirements on a solution to the problem. First, the definition must be *formally correct*, by which he means that it must satisfy the usual rules for constructing definitions—including the rule that the *definiendum* not be defined in terms of (or conceptually dependent upon) any expressions used in the definition. Of course, the definition of truth in L employs standard logical vocabulary—quantifiers, identity, and truth-functional connectives. In section 15 of Tarski (1944), he responds to the objection that since this vocabulary is itself defined in terms of truth, his definition is *not* formally correct. Although his discussion there is not as straightforward as it might have been, his main point is correct. The basic logical vocabulary is not defined but primitive. For truth-functional connectives, this must be so. One can't noncircularly define 'or' by saying *A sentence 'P or Q' is true iff P is true or Q is true*. One can define some truth-functional connectives in terms of others, but since one can't define them all, some must be primitive. The point extends to other logical vocabulary.

Tarski's final requirement is that the truth definition not employ, or depend on, any semantic terms—like *refers to*, *denotes*, or *applies to*. Since they give rise to paradoxes similar to those involving truth, his goal of insulating his formally defined predicate from paradox led him to demand a definition free of semantic primitives. He commented that such a definition "will fulfill what we intuitively expect from every definition; that is it will explain the meaning of the term being defined in terms whose meaning appears to be completely clear and unequivocal. And, moreover, we have then a kind of guarantee that the use of semantic concepts will not involve us in any contradictions."[23]

[23] Tarski (1944 [1952]), section 9, p. 23.

3.4 Dispelling the Illusion

S1. Homophonic instances of *Schema T*rue, like *'snow is white' is true (in a fragment E of English) iff snow is white*, can be known to be true and assertable simply by understanding and reflecting on them.
S2. If the metalanguage E+ of a Tarskian truth definition contains E, and the truth definition in E+ of Tarski's predicate 'T_E' entails homophonic instances of *Schema T*$_E$ in which 'T_E' plays the role of 'true', these instances can also be known to be true and assertable simply by understanding and reflecting on them.
S3. Thus, for each sentence S of E one can establish ⌜'S' is T_E iff 'S' is true in E⌝ simply by understanding 'true in E' and '"T_E'.
S4. Since no empirical information is required to establish this result, ⌜'S' is true in E⌝ and ⌜'S' is T_E⌝ are conceptually equivalent (in effect, synonymous). Each is conceptually equivalent to S.
S5. Similar results can be obtained for cases in which the metalanguage of a Tarskian truth definition does not contain the object language.
S6. So, materially adequate, formally correct definitions of truth predicates capture the ordinary concept of truth when restricted to those languages.

The only step in this argument that is correct is step 2.

The problem with step 1 is illustrated by (9).

9. 'Snow is white' is a true sentence of English iff snow is white.

Suppose I speak English, but don't know that the name 'English' refers to my language. I understand the name and know several things about it—e.g., that it designates a language spoken in England, North America, Australia, and New Zealand. But I don't know that it designates a language I speak. This is possible, just as it is possible for me (i) to understand the name 'Japanese' without knowing it is the language I hear on channel 25, (ii) to understand the name 'Santa Monica' without knowing it is the city in which I am presently located, or (iii) to understand the name 'August' without knowing that it designates the current month. If I am in this situation and don't know that 'English' designates the language I am using when considering (9), then I may not be in a position to know that (9) is both true and assertable.[24]

This result can't entirely be avoided by dropping the explicit reference to English as is done in (10).

10. 'Snow is white' is true iff snow is white.

When we say of a sentence that it is true, what we are really saying is that the proposition it would express, when used in accord with the conventions governing it, is true. Often those are the linguistic conventions of a

[24] See Soames (1999), pp. 238–44.

language, in which case we speak of the sentence being a truth of the language. Although it is possible to say of a sentence merely "It's true," meaning that it expresses a truth in some language or other, usually we have a particular language in mind. In fact, we usually have specific conventions in mind, determining a specific proposition expressed by the sentence.[25] If I were to utter (10) to myself, I would be in an optimally good position to recognize that the predication of *truth* involving the sentence on the left that I *mention* targets the very proposition I express by the sentence on the right that I *use*. Because predications of truth to sentences depend on the connections they bear to propositions, which are never more transparent that the sentence-proposition connection illustrated by this use of (10), this statement of truth conditions is more transparent than the corresponding statement using (9).

The larger point for Tarski is that if sentences are what he thinks they are—syntactically individuated structures of words—then they can be bearers of truth only when the ascriptions of truth to them are relativized to something—to propositions they are used to express, to repeatable act types that are themselves uses of sentences, or to something else. For Tarski, truth is ascribed to sentences relative to the conventions that govern them in one or another public, sharable language, including those used by logicians and mathematicians. How we should understand such relativizations, and which kind of relativization is best—to propositions, to uses, to languages—were not, and did not need to be, on his radar. As the discussion of (9) indicates, however, if *languages* in common use are selected, then bare uses like that in (10), in which the relativization is implicit, will lose some of their transparency and fail to be knowable simply by virtue of understanding them for the same reasons that (9) fails to be knowable simply by understanding it.

Next consider S3. It would be incorrect even if we could establish S1 and S2. If we could do that we could show that ⌜'snow is white' is true in E iff snow is white⌝ and ⌜'snow is white' is T_E iff snow is white⌝ *can be known to be true simply by understanding them*. But we couldn't show that ⌜'snow is white' is T_E iff 'snow is white' is true in E⌝ *can be known to be true simply by understanding it,* because one can understand that sentence without understanding the sentence 'snow is white'.[26] This is significant. We already know we can use the material adequacy of a Tarskian truth definition to establish the *coextensiveness* of 'T_E' with 'true in E'. What is now at issue is a stronger conceptual connection between the two. We can't establish such

[25] The sentence in question may also express other propositions, determined by different conventions associated with an ambiguous word or syntactic construction it contains. Thus in predicating truth of an ambiguous sentence, we usually have specific disambiguating conventions in mind.

[26] To understand the biconditional I must understand the quote-name of the sentence, but that doesn't require understanding the sentence itself.

a connection if we modify S3 to allow antecedent knowledge of crucial properties of the sentences to which the predicates are applied.

Step 4 is also incorrect (on independent grounds). Suppose, for the sake of argument, that merely understanding ⌜'S' is true in E iff S⌝ were sufficient to know it to be true and assertable. This wouldn't establish the conceptual equivalence of ⌜'S' is true in E⌝ and S. For that to hold, the propositions expressed by the two sentences would have to be necessary and a priori consequences of each other, which they aren't. It is a contingent matter which linguistic conventions endow a sentence with meaning. When p is the proposition expressed by S, there will typically be possible conventions that would have rendered S false *had they governed S*, without affecting the truth of p. As a result, the proposition expressed by ⌜'S' is true in E iff S⌝ isn't necessary, and ⌜'S' is true in E⌝ and S are *not* necessary consequences of each other. Moreover, learning the meaning of a sentence requires acquiring empirical evidence about the linguistic conventions governing it. Because of this, there are cases in which understanding S involves having empirical information that provides *justifying evidence* required to warrant accepting the proposition S expresses.[27] Without ruling out the possibility that ⌜'S' is true in E iff S⌝ is such a sentence, one could not establish that it expresses a truth that is knowable a priori (or that its left and right sides are a priori consequences of one another) even if it could be known to be true merely by understanding it.

Step 5 presents different problems. It is instructive to compare *homophonic* instances of *Schema True* like (2a) with *non-homophonic* instances like (11) and (12), which express the same proposition (2a) does.

2a. 'John gave the book to Mary' is true in E iff John gave the book to Mary.
11 'John gave the book to Mary' is true in E iff John gave Mary the book.
12. 'John gave the book to Mary' es verdadero si y solo si le Juan dio el libro a María.

Even if S1 could be established for the homophonic (2a), it could not be established for (11) and (12). Since it is possible to understand the latter without understanding the object-language sentence of which truth is predicated, understanding these instances of *Schema True* doesn't, by itself, warrant taking them to be true or assertable. This is significant, because it is vanishingly rare for Tarskian truth definitions to generate *homophonic* instances of *Schema T*. Very often, the metalanguage doesn't contain the object language, in which case the best one can hope for are instances like (12). Even when it does contain the object language, the relevant instances of *Schema T* are rarely homophonic, and can be made so only with great effort. In these cases there is no hope of establishing that both ⌜'S' is true in E iff P_S⌝ and ⌜'S' is T_E iff P_S⌝ can be known to be true and assertable

[27] For cases of this sort, see Soames (2005b), pp. 56–67.

merely by understanding them—where P_S is the paraphrase of S given in the Tarskian truth definition. Hence, even ignoring the original problem with S1, one can't establish it when the Tarskian truth definition doesn't yield homophonic instances of *Schema T*.

At this point it is useful to recall something else initially said about the homophonic example (2a)—namely that one could acquire the knowledge needed to satisfy the predicate (2b), merely by understanding and reflecting upon (2a).

2b. x knows *that 'John gave the book to Mary' is true$_L$ iff John gave the book to Mary*.

One who accepted S1 might also accept the following reformulation of the argument.

S1*. One who understands a homophonic instance of *Schema True*, like *'John gave the book to Mary' is true (in E) iff John gave the book to Mary*, can thereby come to know the proposition it expresses—i.e., can thereby come know *that 'John gave the book to Mary' is true (in E) iff John gave the book to Mary*, simply by understanding that instance of the schema.

S2*. Let D be a materially adequate definition that entails ⌜*'S' is T_E iff P_S*⌝, where P_S expresses the same proposition as S. When S = 'John gave the book to Mary' and P_S is any sentence that expresses the same proposition as S, anyone who understands that instance can thereby come to know *that 'John gave the book to Mary' is T_E iff John gave the book to Mary*.

S3*. Thus, for each sentence S of E, if S means *that so-and-so*, one can come to know *that S is T_E iff so-and-so* simply by understanding the Tarskian definition of T_E while knowing *that S is true in E iff so-and-so* simply by understanding a homophonic instance of *Schema True*. One can then come to know *that S is T_E iff S is true in E* without further justifying evidence.

S4*. Since one can come to know this without empirical information beyond that needed to understand S, the truth predicates are conceptually equivalent.

There are two old errors here—one in S1* about homophonic instances of *Schema True*, and one in S4* of ignoring how facts about the conventions governing S, and the need for empirical knowledge of them, undermines claims about the conceptual equivalence of S and the claim that S is true. There is also a new error in S3*; its second sentence doesn't follow from its first. One can know that (a) and also know that (b) in each of the following cases, without knowing, or being in any position to know, that (c).

13a. What Pierre said to his friend in Paris is true iff London is pretty.
b. What Pierre said to his London neighbor is true iff London is not pretty.
c. What Pierre said to his friend in Paris is true iff what Pierre said to his London neighbor is not true.

14a. What Mary said is that Carl Hempel was a great man.
b. What Bill said is that Peter Hempel was a great man.
c. What Mary said is true iff what Bill said is true.
15a. What Mary said is that the liquid in the vial is water.
b. What Bill said is that the liquid in the vial is H_2O.
c. What Mary said is true iff what Bill said is true.

The failure to see this point is due to an uncritical acceptance of a principle that, until recent decades, was widely taken for granted in philosophy.

THE TRANSPARENCY OF SAMENESS AND DIFFERENCE OF MEANING

> Anyone who understands a pair of sentences will know (or be in a position to know by reflection) that they mean the same thing iff they do mean the same thing, and will know (or be in a position to know by reflection) that they don't mean the same thing iff they don't.

The most familiar challenge to this principle comes from theories of direct reference. These theories take the semantic contents of sentences that differ only in the substitution of coreferential proper names or simple natural kind terms to be identical—despite the fact that understanding the sentences doesn't, by itself, put one in a position to know that they agree in meaning or truth value.[28] The second challenge comes from the discussion in Kripke (1979) of a monolingual French speaker, who, having seen postcards of London, understands and accepts 'Londres est jolie', while also understanding and rejecting its translation 'London is pretty', after moving across the channel and learning English by immersion. Taken together the challenges show that one can understand a non-homophonic instance ⌜'S' is true iff P_S⌝ of *Schema True*, without realizing that its right side is a paraphrase of S, and without being able to derive it from a homophonic instance ⌜'S' is true iff S⌝.

Salmon (1990) and Reiber (1992) extend the challenges to the transparency principle by showing that agents who understand sentences differing only in the substitution of uncontroversial synonyms—'doctor'/'physician', 'ketchup'/ 'catsup', 'dwelling'/'abode'—may fail to recognize that they agree in truth value. Soames (1986) amplifies the point by showing how even speakers who understand synonyms such as 'fortnight' and 'period of fourteen days', and also understand sentences in which they occur embedded under extensional or intensional operators, may fail to recognize the extensional equivalence of pairs of *attitude ascriptions* they understand in which one of the synonymous terms is substituted for the other.[29]

The errors I have identified in attempts to establish a close conceptual connection between Tarski's truth predicate for L and our ordinary truth

[28] See, e.g., Salmon (1986) and Soames (2002).
[29] See in particular section IX of Soames (1986).

predicate, restricted to L, don't touch the material adequacy or formal correctness of his definitions, or their utility for his meta-mathematical purposes. They do affect the philosophical significance of his definitions, to which I will return in section 5.

4. GIVING THE TRUTH DEFINITION

In order to construct a Tarskian definition of truth for a standard first-order language L of the predicate calculus, one must first give a syntactic description of L that specifies its vocabulary plus the rules for constructing compound expressions from it.[30]

Vocabulary

Nonlogical:

There are finitely many names, function signs (indicating for each the number of its arguments), and predicates (indicating for each the number of its arguments).

Logical:

There are truth-functional connectives—'~', '&', 'v', '→', '↔'—plus '∀', '∃' used to form quantifiers, along with variables 'x_1' 'x_2', etc. Sometimes '=' is included.

Terms

Names and variables are terms; If $t_1 \ldots t_n$ are terms and f is an n-place function sign, the result of combining them is a term.[31] Nothing else is a term.[32]

Formulas

a. A combination of an n-place predicate with n terms is a formula.[33]
b. If A and B are formulas, so are ⌜(~A)⌝, ⌜(A & B)⌝, ⌜(A ∨ B)⌝, ⌜(A→B)⌝, ⌜(A↔B)⌝.
c. If A is a formula and ν is a variable, then so are ⌜(∀νA)⌝ and ⌜(∃νA)⌝.
d. Nothing else is a formula.

Sentences

a. A sentence is a formula containing no free occurrences of any variable.

[30] Throughout the truth definition I use variables of the metalanguage that range over expressions of the object language. For example, in *If A and B are formulas of L, then ⌜(A & B)⌝ is also a formula of L,* 'A' and 'B' are variables ranging over formulas of L. For any assignment to these variables, ⌜*(A & B)*⌝ stands for the expression of L that consists of the left parenthesis, followed by the formula of L assigned to 'A', followed by '&', followed by the formula of L assigned to 'B', followed by the right parenthesis.

[31] I abstract away from the precise syntactic way of combining them, which includes ⌜$f(t_1 \ldots t_n)$⌝ as an option.

[32] In principle, Fregean definite descriptions could also be allowed as terms.

[33] Again, I abstract away from the precise syntax of combination.

b. An occurrence of a variable v in a formula A is free iff it is not within the scope of any occurrence of a quantifier using v.
c. The scope of an occurrence of a quantifier is the quantifier itself plus the smallest (complete) formula immediately following it.[34]

Next we move to the truth definition, which can be broken into three steps. The first consists of definitions of the *application* of an n-place predicate to an n-tuple of objects and the *denotation* of a term relative to an assignment of values to variables—where *assignments are functions from variables to objects in the domain of L* (the class of objects L is used to talk about). More precisely, the defined notions are formal counterparts *application*$_T$ and *denotation*$_T$ of our pretheoretic semantic notions *application* and *denotation*. The relation between our ordinary notions and their Tarskian counterparts can be assessed only after the definitions have been considered.

Once Tarskian substitutes for these ordinary semantic notions are defined, we will be able to define his substitute for the informal notion of a formula (which may contain one or more free variables) *being true of an object or n-tuple of objects*. My Tarskian substitute for this is the notion of *a formula being true*$_T$ *relative to an assignment A of values to variables*. This will be defined by (i) stipulating that the simplest (atomic) formulas—which consist of an n-place predicate plus n terms—are *true*$_T$ relative to A iff the predicate *applies*$_T$ to the *denotations*$_T$ of those terms relative to A, and (ii) extending this result to complex formulas involving truth-functional connectives and quantifiers. The final step in Tarski's truth definition is to define the *truth*$_T$ *of a sentence* in terms of the *truth*$_T$ *of a formula relative to an assignment*.

In giving the sample truth definition, I will depart from Tarski's practice in three ways. First, the notion here called *truth*$_T$ *of a formula relative to an assignment of values to variables in L* is an amalgam of what Tarski calls the *satisfaction of a formula by an infinite sequence of objects* plus an alphabetization of all variables in L, so that the i^{th} variable in the alphabetization can be seen as denoting the object occupying the i^{th} place in a sequence. So, if $F(x_i)$ is a formula in which only x_i has free occurrences, then my statement *F(x$_i$) is true relative to some assignment A of o to x$_i$* expresses what Tarski would express by saying *some sequence S in which object o occupies the ith place satisfies F(x$_i$)*. This difference in formulation is merely terminological.

The second difference is that whereas I give trivial definitions of three concepts used in the definition *truth*$_T$ *relative to an assignment*—the concepts of *a predicate applying*$_T$ *to an object (or objects)*, *a name denoting*$_T$ *an object*, and *a function sign denoting*$_T$ *a function*—Tarski folds my trivial definitions into clauses of the single definition of the *satisfaction of a formula by a sequence*. This difference is cosmetic.

[34] These definitions rely on complete specifications of formulas, including parentheses. When no ambiguity results, I drop unneeded parentheses to reduce clutter.

Finally, instead of selecting a mathematical object language like the calculus of classes to provide the sentences to which the defined Tarskian truth predicate is to apply, I will take the object language to be one we imagine being used to talk about ordinary things, including people and places. I do this to make the discussion, which is already abstract enough, more intuitive to some readers, but also to facilitate later comparisons between ascriptions of our ordinary truth predicate to sentences and ascriptions of Tarski's predicate. One point to note at the outset is that the domain of the object language L may include objects not denoted by any names or (closed) singular terms of L, even though they are among the things to which some of L's predicates apply, while also being in the range of L's quantifiers—in much the way that the English predicates 'is a grain of sand' and 'is an electron' apply, respectively, to each grain of sand and to each electron, which in turn are in the range of English quantifiers.

The first step in our Tarskian truth definition consists of explicit definitions of *the denotation$_T$ of a name* and *the denotation$_T$ of a function sign*.

The Denotation$_T$ of a Name

For all names n of L and objects o (of the domain), n denotes$_T$ o iff n = the name 'a' and a = Albert, or n = the name 'Brian' and o is Brian, or . . . (and so on, until we have a clause stipulating the denotation$_T$ of each name).

The Denotation$_T$ of a Function Sign

For all n-place function signs h of L, and n-place functions f from n-tuples of the domain into the domain, h denotes$_T$ f iff

(i) h = the one-place function sign 'g_{1a}' and f is the one-place function that assigns the spouse of its argument as value, if it has a unique spouse, and otherwise assigns the argument itself as value, or h = the one-place function sign 'g_{1b}' and f is the one-place function that assigns the argument's birthplace as value, if it had a birthplace, and otherwise assigns the argument itself as value . . . (and so on until we have a clause stipulating the denotation$_T$ of each one-place function sign); or

(ii) h = two-place function sign 'g_{2a}' and f is the two-place function that assigns the average mass as value to a pair of arguments, if they have mass, and zero if neither do, or h = the two-place function sign 'g_{2b}' and f is the two-place function that assigns their most recent common ancestor as value to a pair of arguments, if they have a common ancestor and zero if they don't . . . (and so on until we have a clause stipulating the denotation$_T$ of each two-place function sign).

(iii) Similar clauses for all n-place function signs for arbitrary n are provided until all function signs in L are exhausted.

These two definitions provide concepts needed in the *inductive definition* of the denotation$_T$ of a term relative to an assignment of values to variables.

INDUCTIVE DENOTATION$_T$ OF A TERM RELATIVE TO AN ASSIGNMENT

(i) For all variables ν of L, objects o of the domain, and assignments A, ν denotes$_T$ o relative to A iff the value of A at the argument ν is o.
(ii) For all names n of L, objects o, and assignments A, n denotes$_T$ o relative to A iff n denotes$_T$ o.
(iii) For all terms T of L that consist of an n-place function sign h plus terms $t_1 \ldots t_n$, objects o, and assignments A, the denotation$_T$ of T relative to an assignment A = o iff h denotes$_T$ a function f that assigns o as value to the n-tuple of objects $o_1 \ldots o_n$ denoted$_T$ by $t_1 \ldots t_n$ relative to A.

Several points are worth noting. First, the relativization of the denotation$_T$ of a term to assignments is needed for complex terms and the variables they contain. These assignments will allow us to reduce the truth$_T$ of quantified sentences to the truth$_T$ or falsity$_T$ of simpler formulas by allowing variables and terms containing variables to "temporarily denote$_T$" selected objects in the domain. This mechanism of allowing each variable ν to function as a "temporary name" of an object, and then computing the denotation$_T$ or truth$_T$ of compound expressions containing ν, will enable us to define the truth$_T$ of quantified sentences in terms of the truth$_T$ or falsity$_T$ of their instances, even when the instances aren't *sentences*.

Second, despite the utility of relativizing the denotation$_T$ of a term to assignments of values to variables, clause (ii) for names ensures that the denotations$_T$ of names are fixed independently of assignments, and so do not vary from one to the next. The only reason for mentioning assignments in (ii) is to have a single concept—*the denotation$_T$ of a term relative to an assignment*—that applies to terms of all types.

Third, clauses (ii) and (iii) are dependent on the earlier definition of *denotation*$_T$ of names and function signs that are simply *lists* of individual stipulations of the denotations$_T$ of those expressions, dressed up in the form of universally quantified biconditionals to count as *formally correct definitions*. The question isn't whether these list-like statements are genuine *definitions*; they are. The question is whether the notion, *denotation$_T$*, they are used to define is a genuine *semantic* counterpart of our pretheoretic semantic notion of *denotation*. Just as there is a real question of whether the notion defined by the trivial, formally correct, universally quantified biconditional (7) is a genuine semantic counterpart to our ordinary notion of *truth*, restricted to the language with only ten sentences, so there is a real issue whether Tarski-denotation, *denotation$_T$*, is anything more than coextensive over L with the genuine semantic notion of *denotation*.

This final point also applies to the list-like definition of what it is for *predicates to apply* to objects.

THE APPLICATION$_T$ OF PREDICATES

a. For all one-place predicates P of L, P applies$_T$ to an object (in the domain) iff P = 'F' and o is female, or P = 'M' and o is male, or . . . (and so on until we have a clause for each one-place predicate of L).

b. For all two-place predicates P of L, P applies$_T$ to a pair of objects (from the domain) iff P = 'L' and o_1 loves o_2, or P = 'H' and o_1 hates o_2, or . . . (and so on until we have a clause for each two-place predicate of L).
c. Similar clauses for all n-place predicates for arbitrary n are provided until all predicate symbols in L are exhausted.

These definitions of *denotation$_T$* and *application$_T$* provide the basis for the *inductive definition* of the *truth$_T$ of a formula relative to an assignment*.[35]

INDUCTIVE DEFINITION OF TRUTH$_T$ RELATIVE TO AN ASSIGNMENT

a. An atomic formula $\ulcorner P\ t_1 \ldots t_n \urcorner$ consisting of an n-place predicate P followed by n terms $t_1 \ldots t_n$ is true$_T$ relative to an assignment A of values to variables iff P *applies$_T$* to the n-tuple of objects $o_1 \ldots o_n$ denoted$_T$ by the terms relative to A.
b$_\sim$. A formula $\ulcorner \sim \Phi \urcorner$ is true$_T$ relative to an assignment A iff Φ is not true$_T$ relative to A.
b$_\&$. A formula $\ulcorner \Phi\ \& \Psi \urcorner$ is true$_T$ relative to an assignment A iff both Φ and Ψ are true$_T$ relative to A.
b*. Similar clauses are provided for other truth-functional connectives, including '$\vee$', '$\rightarrow$', '$\leftrightarrow$'.
c$\exists$. A formula $\ulcorner \exists \nu\ \Phi(\nu) \urcorner$ is true$_T$ relative to an assignment A iff there is an object o and assignment A* that assigns o as value of ν *that is identical with A or that differs from A only in the value it assigns to* ν, and the formula $\Phi(\nu)$ which arises from $\ulcorner \exists \nu\ \Phi(\nu) \urcorner$ by erasing $\ulcorner \exists \nu \urcorner$ is true$_T$ relative to A*.
c$\forall$. A formula $\ulcorner (\forall \nu\ \Phi(\nu)) \urcorner$ is true$_T$ relative to an assignment A iff for every object o and assignment A* *that is identical with A, or that differs from A only in assigning o to* ν, the formula $\Phi(\nu)$ which arises from $\ulcorner \forall \nu\ \Phi(\nu) \urcorner$ by erasing $\ulcorner \forall \nu \urcorner$ is true$_T$ relative to A*.

Finally, we must define the unrelativized notion of truth$_T$ applying to *sentences* using the relativized notion defined for formulas. There are two natural choices for doing this. We could say that a sentence is true$_T$ iff it is true$_T$ relative to some assignment, or we could say that it is true$_T$ iff it is true$_T$ relative to all assignments. It makes no difference which we select because any *sentence* that is true$_T$ relative to one assignment is true$_T$ relative to all assignments.

This can be shown by verifying (16).

16. If a formula F of L is true$_T$ (not true$_T$) relative to an assignment A, then F is true$_T$ (not true$_T$) relative to all assignments that differ from A only in what they assign variables that have no free occurrences in A.

Since sentences have no free occurrences of variables, it follows from (16) that a sentence that is true$_T$ relative to one assignment is true$_T$ relative to all assignments.

[35] Outer parentheses on a formula are dropped in what follows when no ambiguity results.

To establish (16), we reason as follows: First, the truth of (16) is transparent for atomic formulas (complexity 0). Next, suppose (16) holds for all formulas of complexity n—where ⌜~Φ⌝, ⌜∃ν Φ(ν)⌝, and ⌜∀ν Φ(ν)⌝ have complexity n+1 when Φ(ν) has complexity n, and where ⌜Φ & Ψ⌝ has the complexity 1 plus the complexity of its most complex conjunct (similarly for disjunctions, conditionals, and biconditionals). Given the supposition that (16) holds for all formulas of complexity n or less, we let Θ be of complexity n+1 and proceed to show that (16) must hold for Θ. We begin by noting that the truth_T, or falsity_T, of a truth-functional compound—⌜~Φ⌝, ⌜Φ & Ψ⌝, ⌜Φ ∨ Ψ⌝, ⌜Φ → Ψ⌝, or ⌜Φ ↔ Ψ⌝—relative to an assignment A is determined by the truth_T, or falsity_T, of its truth-functional constituents—Φ, Ψ—relative to A. So if Θ is such a compound, and (16) holds of its constituents Φ and Ψ, then, (16) must hold for Θ. Next, we let Θ be ⌜∃ν Φ(ν)⌝. It then follows from clause (c∃) above that if Θ is true_T relative to some assignment A, then there is an assignment A* that is either identical with A or differs from A only in what it assigns ν, at which Φ(ν) is true_T. Since, by supposition, (16) holds for Φ(ν), Φ(ν) is true_T relative to all assignments A*' that agree with A* on assignments to variables that no have free occurrences in Φ(ν). Now let A' be any assignment that agrees with A on all variables that have free occurrences in Θ. Invoking clause c∃ again, it follows that for each such A' there will be an A*' that is either identical with A' or differs from A' only in what it assigns ν, which is such that Φ(ν) is true_T relative to A*'. So Θ is true_T relative to all A' that either don't differ from A at all, or differ from it only in what is assigned to v. This means that (16) holds for Θ. Since we get the same result with ⌜∀ν F(ν)⌝, we verify that (16) is true, and hence that any *sentence that is true_T relative to at least one assignment is true_T relative to all assignments*. Thus, we may define Tarski-truth for L as follows.

Definition of Sentential Truth_T for L

For all sentences S of L, S is true_T in L iff S is true_T relative to some assignment of values to variables of L (which holds iff S is true relative to all assignments).

The only aspect of the definition *truth_T relative to an assignment* that may not be immediately obvious is clause (c∃). One might wonder why the truth_T of ⌜∃ν Φ(ν)⌝ relative to A is defined in terms of the truth_T of Φ(ν) relative to *some assignment that may differ from A*, rather than A itself. Why isn't the clause (c$\exists_*$) rather than (c∃)?

c$\exists_*$. A formula ⌜∃ν Φ(ν)⌝ is true_T relative to an assignment A iff the formula Φ(ν) which arises from ⌜∃ν Φ(ν)⌝ by erasing ⌜∃ν⌝ is true_T relative to A.

The answer is that adopting (c$\exists_*$) would subvert the material adequacy of the definition of truth_T for sentences of L. If (c$\exists_*$) were substituted for (c∃), (16) would be falsified, and we would have to decide whether to define Tarskian sentential truth_T as *truth_T relative to some assignment* or as *truth_T relative to all assignments*. On the former option, (17a) is a consequence of the definition of truth_T and (17b) fails to be such a consequence.

17a. ⌜~∃x(Mx ∨ Fx)⌝ is true_T in L iff at least one object (in the domain) is neither male nor female.

b. ⌜~∃x(Mx ∨ Fx)⌝ is true_T in L iff no object (in the domain) is either male nor female.

On the latter option, (18a) is a consequence of the definition of truth_T and (18b) fails to be a consequence.

18a. ⌜∃x (Mx ∨ Fx)⌝ is true_T in L iff every object (in the domain) is male or female.

b. ⌜∃x (Mx ∨ Fx)⌝ is true_T in L iff at least one object (in the domain) is male or female.

In both (17b) and (18b) the metalanguage sentences appearing to the right of 'iff' *are* paraphrases of the sentences of L mentioned on the left, while in (17a) and (18a) the metalanguage sentences to the right of 'iff' *are not* paraphrases of the sentences of L mentioned on the left. Let us suppose, for the sake of argument, that the domain of L includes at least one human being who is either male or female and one inanimate object which is neither. Then the sentence of L mentioned on the left of the biconditionals in (17) is *false*, despite being $true_T$, while the sentence of L on the left of the biconditionals in (18) is true, despite being $false_T$. Since to derive these results would be to demonstrate that one's definition is *not* materially adequate and that the notion $truth_T$ one defined is *not* coextensive with *true sentence of L,* Tarski could not have adopted clause (c∃$_*$).

The material adequacy of the original definition, using (c∃), is illustrated by the following derivation, where for simplicity we may take the domain to consist only of human beings and the two-place predicate 'L' to apply to a pair of individuals iff the first loves the second.

S1. '$\exists x_1$ (~($\exists x_2$ (Lx_1x_2)))' is true_T iff it is true_T relative to an assignment A of values to variables. (From the *Definition of Sentential* $Truth_T$ *for L*)

S2. '$\exists x_1$ (~($\exists x_2$ (Lx_1x_2)))' is true_T relative to an assignment A iff there is an object o_1 and an assignment A* that assigns o_1 as value of 'x_1' that is either identical with A or differs from A only in what it assigns 'x_1', and '~($\exists x_2$ (Lx_1x_2))' is true_T relative to A*. (From clause (c∃) of the *Definition of* $Truth_T$ *Relative to an Assignment*)

S3. '~($\exists x_2$ (Lx_1x_2))' is true_T relative to A* iff '$\exists x_2$ (Lx_1x_2)' is not true_T relative to A* (where A* is either identical with A or differs from A only in assigning o_1 to 'x_1'). (From clause (b$_\sim$) of the definition of $Truth_T$ *Relative to an Assignment*)

S4. '$\exists x_2$ (Lx_1x_2)' is not true_T relative to A* iff there is no object o_2 and assignment A*' such that (i) A*' assigns o_2 to 'x_2', (ii) A*' is either identical with A* or differs from A* only in what it assigns to 'x_2', and (iii) '(Lx_1x_2)' is true_T relative to A*'. (From clause (c∃) again)

S5. '(Lx_1x_2)' is true_T relative to A*' iff the predicate 'L' applies_T to the pair consisting of the denotation_T of 'x_1' relative to A*' followed by the denotation_T

of 'x_2' relative to A*' (where A*' is either identical with A* or differs from it only in assigning o_2 to 'x_2'). (From clause (a))

S6. 'L' applies$_T$ to the pair consisting of the denotation$_T$ of 'x_1' relative to A*' followed by the denotation$_T$ of 'x_2' relative to A*' iff o_1 (the individual A*' assigns 'x_1') loves o_2 (the individual A*' assigns 'x_2'). (From the *Definition of Application$_T$ of Predicates* and clause (a) of the *Definition of the Denotation$_T$ of a Term Relative to an Assignment*)

S7. '(Lx_1x_2)' is true$_T$ relative to A*' iff o_1 (the individual A*' assigns 'x_1') loves o_2 (the individual A*' assigns 'x_2'). (From S5 and S6)

S8. '$\exists x_2\ (Lx_1x_2)$' is not true$_T$ relative to A* iff there is no object o_2 and assignment A*' such that (i) A*' assigns o_2 to 'x_2', (ii) A*' is either identical with A* or differs from A* only in what it assigns to 'x_2', and (iii) the individual A*' assigns 'x_1' loves the individual o_2 that A*' assigns 'x_2'. (From S4 and S7)

S9. There is no object o_2 and assignment A*' such that (i) A*' assigns o_2 to 'x_2', (ii) A*' is either identical with A* or differs from A* only in what it assigns to 'x_2', and (iii) the individual A*' assigns 'x_1' loves the individual o_2 that A*' assigns 'x_2' iff there is no individual who is loved by the individual that both A* and A*' assign 'x_1'. (From (i) the stipulation that for any individual and any variable, there are assignments of that individual to that variable and (ii) the stipulation that A*' is either identical with A* or differs from it only in what it assigns 'x_2')

S10. '$\exists x_2\ (Lx_1x_2)$' is not true$_T$ relative to A* iff there is no individual who is loved by the individual A* assigns 'x_1'. (From S8 and S9)

S11. '$\sim(\exists x_2\ (Lx_1x_2))$' is true$_T$ relative to A* iff there is no individual who is loved by the individual A* assigns 'x_1'. (From S3 and S10)

S12. '$\exists x_1\ (\sim(\exists x_2\ (Lx_1x_2)))$' is true$_T$ relative to an assignment A iff there is an object o_1 and an assignment A* that assigns o_1 as value of 'x_1' that is either identical with A or differs from A only in what it assigns 'x_1', and no individual is loved by o_1. (From S2 and S11)

S13. '$\exists x_1\ (\sim(\exists x_2\ (Lx_1x_2)))$' is true$_T$ relative to an assignment A iff there is some individual who loves no individual. (From S12)

S14. '$\exists x_1\ (\sim(\exists x_2\ (Lx_1x_2)))$' is true$_T$ iff there is some individual who loves no individual. (From S12 and the *Definition of Sentential Truth$_T$*)

In addition to illustrating the material adequacy of the *Definition of Sentential Truth$_T$ for L*, this derivation also illustrates the importance of the italicized restriction below on assignments A* in clause (c∃) of the *Definition of Truth$_T$ Relative to an Assignment*.

c∃. A formula $\ulcorner\exists\nu\ \Phi(\nu)\urcorner$ is true$_T$ relative to an assignment A iff there is an object o and assignment A* that assigns o as value of ν *that is identical with A or that differs from A only in the value it assigns to* ν, and the formula $\Phi(\nu)$ which arises from $\ulcorner\exists\nu\ \Phi(\nu)\urcorner$ by erasing $\ulcorner\exists\nu\urcorner$ is true$_T$ relative to A*.

If this restriction were deleted, and (c$\exists_{**}$) were substituted for (c∃), then S4 would be replaced by S4*, and S14* would be derived instead of S14.

c∃** A formula ⌜∃ν Φ(ν)⌝ is true_T relative to an assignment A iff there is an object o and assignment A* that assigns o as value of ν, and the formula Φ(ν) that arises from ⌜∃ν Φ(ν)⌝ by erasing ⌜∃ν⌝ is true_T relative to A*.

S4*. '$\exists x_2\ (Lx_1x_2)$' is not true_T relative to A* iff there is no object o and assignment A*' that assigns o as a value to 'x_2' (and also assigns values to all other variables of L, including 'x_1') such that '(Lx_1x_2)' is true_T relative to A*'.

S14*. '$\exists x_1\ (\sim(\exists x_2\ (Lx_1x_2)))$' is true_T iff no one loves anyone.

Thus the restriction on the assignment A* in clause (c∃) is needed to ensure the material adequacy of the definition.

Despite the material adequacy of the definition of *the truth$_T$ of a sentence*, we haven't yet shown it to be formally correct in Tarski's sense because, as given, it depends on a concept *the truth$_T$ of a formula relative to an assignment* for which we have so far given only an inductive definition. Tarski wanted something stronger. He required explicit definitions of semantic notions in the form of universally quantified biconditionals in which the *definiendum* (the semantic term occurring on the left) is defined by a formula occurring on the right that is free, not only of the *definiendum*, but also of any other semantic terms that aren't themselves *explicitly* and noncircularly defined. The end result was to be a definition resulting in a complex formula of the metalanguage, free of all semantic terms, that is capable of replacing the *definiendum*, *true$_T$*, in every sentence of the metalanguage, without loss of expressive power.

There is a standard technique that can be used to transform the inductive definitions we have offered into explicit definitions. Inductive definitions employ the term being defined—e.g., *denotation$_T$ relative to an assignment* and *truth$_T$ relative to an assignment*—in clauses that specify its application to new cases by virtue of its application to previously defined cases. To transform such a definition into an explicit definition, one trades such occurrences of the term being defined for occurrences of a variable ranging over sets, and rewrites the clauses to specify set-membership conditions for new cases in terms of set membership for previous cases. The transformation is completed by putting the definition in the form of a universally quantified biconditional with explicit quantification over sets introduced in the *definiens* (the defining clause).

Here is the transformation of our first inductive definition of a semantic term.

THE DENOTATION$_T$ OF A TERM RELATIVE TO AN ASSIGNMENT OF VALUES TO VARIABLES

For all terms T of L, objects o (of the domain), and assignments A of values to variables, *T denotes$_T$ o relative to A* if and only if

there is a set D of which <T, o, A> is a member, and for all expressions e, objects o', and assignments A', <e, o', A'> is a member of D iff

(i) e is a variable ν to which A' assigns o' as value; or

(ii) for some name n, e = n, *n denotes$_T$ o'*; or

(iii) for some n-place function sign h and terms $t_1 \ldots t_n$, e = ⌜h($t_1 \ldots t_n$)⌝, *h denotes$_T$ a function f* that assigns o' as value to the n-tuple of objects $o_1 \ldots o_n$ such that for each i <t_i, o_I, A'> is a member of D.

This is an explicit definition in which the formula following 'if and only if' defines, and can be used to replace, the *definiendum*—which is the formula *T denotes$_T$ o relative to A* that precedes 'if and only if'. Although this formula itself contains two semantic terms—*n denotes$_T$ o'* and *h denotes$_T$ a function f*—they have already been given explicit, formally correct definitions, and so can be replaced without loss of expressive power by metalanguage formulas free of semantic terms. In this way, the concept *denotation$_T$ of a term relative to an assignment* can be expressed in the metalanguage by a formula in which no semantic terms occur.

To satisfy Tarski's condition of formal correctness, we need only to transform the previous definition of *truth$_T$ relative to an assignment* into an explicit definition.

THE TRUTH$_T$ OF A FORMULA RELATIVE TO AN ASSIGNMENT

For all formulas F of L and assignments A of values to variables, F is true$_T$ relative to A if and only if there is a set T of which <F, A> is a member and for all formulas G and assignments A', <G, A'> is a member of T iff

(i) G = ⌜P $t_1 \ldots t_n$⌝ for some predicate P and terms $t_1 \ldots t_n$, P *applies$_T$* to $o_1 \ldots o_n$, which are *denoted$_T$ by $t_1 \ldots t_n$ relative to A'*; or

(ii) G = ⌜~Φ⌝ for some formula Φ, and <Φ, A'> is not a member of T; or

(iii) G = ⌜Φ & Ψ⌝ for some formulas Φ and Ψ, and both <Φ, A'> and <Ψ, A'> are members of T, or G = . . . (similarly for 'v', '→', and '↔'); or

(iv) G = ⌜∃v Φ(v)⌝ for some formula Φ(v), and there is an object o and assignment A'* such that A'* assigns o as value of v, and A'* is either identical with A' or differs from A' only in what it assigns to v, and <Φ(v), A'*> is a member of T; or

(v) G = ⌜∀v Φ(v)⌝ for some formula Φ(v), and, for every object o and assignment A'* that is identical with A', or that differs from A' only in assigning o to v, <Φ(v), A'*> is a member of T.

As before, this is an explicit definition in which the formula following 'if and only if' defines, and can be used to replace, the *definiendum*—which is the formula *F is true$_T$ relative to A* that precedes 'if and only if'. Also as before, this formula contains two semantic terms—in this case *applies$_T$* and *denotation relative to an assignment*. Since they have already been given formally correct definitions, they can be replaced, without loss of expressive power, by semantics-free formulas of the metalanguage. Consequently, Tarski's *definiendum—the truth of a sentence of L*—can also be so replaced. With this, he succeeded in his goal of providing a concept that is both coextensive with *truth* over L and expressed by a formula in which no semantic terms occur.

There is only one formal limitation, which Tarski clearly recognized. The technique for turning an inductive definition into an explicit definition will work only when the existence of the set required by the explicit definition is guaranteed. One will have that guarantee whenever the domain of the object language is a set and the metalanguage contains quantifiers that range over arbitrary n-tuples of the elements of that set. But when the quantifiers of the object language range over all sets—with the consequence that the domain of the language is not a set—no explicit Tarskian definitions of *denotation*$_T$ or *truth*$_T$ *relative to an assignment* are possible.

To see this, let L be such an object language, and imagine trying to construct an explicit definition of *denotation*$_T$ *relative to an assignment* of the sort just illustrated. If there were a set D of the kind required above, it would be in the domain of L, and would therefore be assigned as value of a variable by some assignments. Our definition would then let us derive (19a), from which we would get the necessarily false (19b).[36]

19a. D is a member of the domain of L iff some assignment A assigns D as value of a variable ν as argument.
b. <D, ν, A> is a member of D.

For this reason, there is no set D of the sort required by an explicit definition of *denotation*$_T$ relative to an assignment, for a language the quantifiers of which range over all sets. Of course, that doesn't mean that we can't speak of terms of such a language denoting things relative to assignments. For although no explicit definition is possible, we could still employ an inductive definition—thought of as a set of axioms that fix the application of the concept inductively "defined."[37]

5. THE SEARCH FOR AN ANALYSIS OF TRUTH

5.1. What Is an Analysis?

Many philosophers have thought that Tarski's definition is a philosophically revealing analysis of the notion of a *true sentence*, as it has classically been understood. They have supposed this even though they have realized

[36] I assume here that n-tuples are set-theoretic constructions, and that no set can be a member of itself, or a member of a member of itself, and so on.

[37] In sections 2 and 3 of Tarski (1935), the definition of truth$_T$ is given for the first-order calculus of classes. In section 4, Tarski extends the techniques of the definition to what he calls *languages of finite order*, which are type-theoretic languages in which nth-order quantification may occur, for any n. In section 5, he considers type-theoretic *languages of infinite order,* for which neither explicit truth definitions nor standard inductive definitions can be given. In this case Tarski simply takes the extension of the concept *true*$_T$ to be fixed by the infinite class of "partial definitions" of the form *X is true*$_T$ *iff P*—where 'P' is replaced by a metalanguage paraphrase of a sentence S of the object language, and 'X' is replaced by a transparent Tarskian name of S.

that his truth predicate $true_T$ for sentences of L doesn't mean the same as the English predicate *true*, when applied to sentences. Whereas $true_T$ applies only to sentences of L—which never contain $true_T$—*true* can be predicated of arbitrary sentences of English (including those containing *true*), as well as of sentences of other languages. Just as we can say, in English, that an English sentence is true, we can also say that a Japanese sentence is true, and even that the first sentence uttered in the twenty-fifth century, or its negation, will be true (without knowing what language will be used). This suggests that the English truth predicate of sentences is, unlike Tarski's, an unrestricted relational predicate *true in L*, for variable L.

How, given these differences, could anyone take Tarski's definition to be an analysis of our ordinary notion? In answering this question, one must remember that according to Tarski our ordinary notion of truth is defective precisely because its unrestrictedness generates paradox. By contrast, it is maintained, Tarski's restricted truth predicate eliminates this defect while preserving the important and useful features of our ordinary notion of truth. On this view, Tarski specifies, not how *true* is actually understood, but how it ought to be understood, if it is to function in our logical, mathematical, and scientific theories in the ways that are normally intended.

Analyses of this kind are sometimes called *explications*. An explication of a pretheoretically understood concept C may consist in the definition of a related concept C* that (i) applies to clear and central instances of C, (ii) is precise and well defined, (iii) is free of difficulties and obscurities that plague C, and (iv) is capable of performing the function of C in all theoretical contexts in which some such notion is required. The claim that Tarski's definition of truth is an analysis of truth is best understood as the claim that the concept $truth_T$ it defines satisfies these criteria.

If we focus on the object languages for which he provided truth definitions, we see that the material adequacy and formal correctness of his definition ensures that criteria (i) and (ii) are met. Criteria (iii) and (iv) raise more interesting issues. To evaluate his purported analysis against them, we must determine what difficulties and obscurities in the ordinary notion of truth it eliminates, and also how well $truth_T$ performs the theoretical functions one can reasonably demand of a truth predicate.

For Tarski, the Liar was the chief difficulty posed by the ordinary notion of truth. But, as sections 5 and 6 of chapter 7 illustrated, several leading philosophers of the period—including Carnap, Neurath, Hempel, and Reichenbach—were skeptics about truth for independent epistemological and metaphysical reasons. It is easy to see why Tarski's definition was historically effective in sweeping away the bases of such skepticism. He showed how to explicitly define truth predicates for certain languages L, presumed to be adequate for science and mathematics, using only notions already expressible in L plus descriptive syntax and elementary set theory. So, if syntax, set theory, and L are all unparadoxical and philosophically unproblematic, then adding Tarski's predicate $true_T$ to a metalanguage

containing all three can't possibly lead to paradox, or to any philosophically objectionable consequences. For any sentence S of L, ⌜'S' is $true_T$⌝ is provably equivalent, in the presence of descriptive syntax and elementary set theory, to S. So, if prior to Tarski one had been inclined toward truth-skepticism, without perhaps seeing how one could entirely do without it, then Tarski's definition might well have seemed to provide a philosophically liberating analysis of what had previously appeared to be a problematic notion.

The final criterion for assessing whether Tarski's definition is an explication of truth is theoretical fruitfulness. Truth is important, and arguably indispensible, for many metatheoretical investigations. Often, we want to know whether all the claims of a given theory are true, whether there are truths it doesn't capture, and whether other theories do better in telling the truth about a specific domain than it does. It would be hard even to formulate these questions without a notion of truth. We also want to know precisely when the truth of a set of sentences *guarantees* the truth of other related sentences. The success of Tarski's characterization of truth is due in large part to its utility in metatheoretical investigations of these kinds.

5.2. The Theoretical Fruitfulness of Tarski's Definition

Tarski's definition not only defines a predicate $true_T$ that applies to all and only the truths of L; it also links the $truth_T$ or $falsity_T$ of a sentence S to the $denotations_T$ (relative to assignments) of the singular terms that occur in S, the $application_T$ of predicates in S to objects, and the $truth_T$ or $falsity_T$ (relative to assignments) of the formulas S contains. Because it does, the definition provides tools sufficient to prove important metatheorems like (20) about the relationship between different object-language sentences (for a certain class of languages).

20. For any singular terms t_1 and t_2 of L, if ⌜$t_1 = t_2$⌝ is $true_T$ and if s and s* are sentences of L that differ only in the substitution of one of these terms for one or more occurrences of the other, then s is $true_T$ iff s* is $true_T$.

Because the Tarskian substitutes for pretheoretic notions of denotation, application, and truth show the same the structural sensitivity and interdependence that the pretheoretic notions do, they can play the roles of those notions in establishing significant results about the object language. Also, given a system of proof for sentences of L, one can give answers to questions like (21)–(23) that match those one would get employing our pretheoretic semantic notions.

21. Is every sentence of L that is provable $true_T$?
22. Is every sentence of L that is $true_T$ provable in the system?
23. If it is not the case that all and only the $truths_T$ are provable in the system, is there any other system in which they are all provable?

Tarski's truth definition also laid the foundation for modern model theory and its use in defining logical truth and consequence. A logical truth is, roughly, a sentence that is Tarski-true, no matter what nonempty domain of quantification is chosen or what denotations assigned to the nonlogical vocabulary. A sentence q is a logical consequence of a set of sentences p iff for every choice of a nonempty domain of quantification plus denotations of the nonlogical vocabulary, q is Tarski-true, if p is.

It is common to take a model M to be a pair consisting of a nonempty domain D_M and a function V_M that assigns members of D_M to names, functions from n-tuples of members of D_M to members of D_M to n-place function signs, and sets of n-tuples of D_M to n-place predicates. *Truth in M* for variable M is abstracted from Tarski's truth definition by first stipulating that variables are assigned members of D_M and each name and function symbol in M denotes the value assigned to it by V_M, and then relativizing *the denotation of a term relative to an assignment* and *the truth of a formula relative to an assignment* to M. The truth of a sentence in M is its truth in M relative to all assignments.

THE DENOTATION IN M OF A TERM RELATIVE TO AN ASSIGNMENT

The denotation in M of a name n relative to an assignment A is the object that V_M assigns to n. The denotation in M of an n-place function sign h relative to an assignment A is the n-place function f_h that V_M assigns to h. The denotation in M of $\ulcorner h(t_1 \ldots t_n)\urcorner$ relative to an assignment A is the value that f_h assigns to the n-tuple of objects $o_1 \ldots o_n$ denoted in M by $t_1 \ldots t_n$ relative to A.

THE TRUTH IN M OF A FORMULA RELATIVE TO AN ASSIGNMENT

An atomic formula $\ulcorner Pt_1 \ldots t_n\urcorner$ is true in M relative to an assignment A iff P *applies* to the n-tuple of denotations $o_1 \ldots o_n$ of $t_1 \ldots t_n$ in M relative to A *iff*$_{def}$ $<o_1 \ldots o_n>$ is a member of the set that V_M assigns to P.

$\ulcorner\sim\Phi\urcorner$ is true in M relative to an assignment A iff Φ is not true in M relative to A.

$\ulcorner\Phi \,\&\, \Psi\urcorner$ is true in M relative to A iff Φ and Ψ are both true in M relative to A.

$\ulcorner\Phi \vee \Psi\urcorner$ is true in M relative to A iff either Φ or Ψ (or both) are true in M relative to A.

$\ulcorner\Phi \rightarrow \Psi\urcorner$ is true in M relative to A iff either Φ is false or Ψ is true in M relative to A.

$\ulcorner\Phi \leftrightarrow \Psi\urcorner$ is true in M relative to A iff both are true or both are false in M relative to A.

$\ulcorner\exists\nu\, \Phi(\nu)\urcorner$ is true in M relative to A iff there is an object o in D_M and assignment A* that assigns o to ν that is identical with A or that differs from A only in what it assigns to ν, and $\Phi(\nu)$ is true in M, relative to A*.

$\ulcorner\forall\nu\, \Phi(\nu)\urcorner$ is true in M relative to A iff for every object o in D_M and assignment A* that is either identical to A or that differs from A only in assigning o to ν, $\Phi(\nu)$ is true in M, relative to A*.

A sentence S is true in a model M iff S is true in M, relative to all assignments of objects of the domain of M to variables.

It is now standard to define logical truth and logical consequence in terms of this notion of truth in a model—logical truths being true in all models, and logical consequences of a set of sentences being true in all models in which the sentences in the set are true. The intuitive rationale for defining the logical consequences of sentences in terms of a certain kind of guaranteed truth preservation—and for defining the related notions of the necessary and a priori consequences of propositions in terms of related kinds of guaranteed truth preservation—is that doing so provides us with effective ways of tracking and systematizing important types of argumentative commitments. Such definitions serve this interest because we take it for granted that the acceptance of S (and the proposition it expresses) carries with it a commitment to the truth of S (and the proposition it expresses), and *vice versa*. This feature of our ordinary concept of truth is central to our practice of using truth-defined consequence relations to track argumentative commitments.

What are these commitments? Consider a simple case. Mary says, "Sam is fat and John is happy." Since doing this commits her to the truth of 'John is happy', and so to John's being happy, for her to then say "John isn't happy" would be for her to incoherently assert and deny the same thing. Avoiding this sort of incoherence is a basic argumentative commitment. We can express this point using the notion of truth. In assertively uttering the conjunctive sentence (and asserting the proposition it expresses), Mary is committed to its truth, and thereby to the truth of 'John is happy' (and the proposition it expresses). So if she were then to say "John isn't happy," she would be committed to the untruth of 'John is happy' in addition to being committed to its truth. In this scenario, Mary can correctly be described in two ways: (i) as incoherently saying of John, *he's happy*, and *he's not*, and (ii) as incoherently saying of 'John is happy' (and the proposition it expresses), *it's true,* and *it's not*. These are two sides of the same coin.

This illustrates how ascent to truth provides a way of generalizing and systematizing our argumentative commitments. The utility of the truth predicate in studying the basic forms of such commitment lies in its role in abstracting away from particular predications and argument forms, and bringing them under a small set of general headings including *logical consequence, logical inconsistency*, and so on. This is the rock on which definitions of logical notions in terms of guarantees of truth, or untruth, stand. It is also the reason such definitions are philosophically more fundamental than any notion of provable sentence defined by particular axioms and formal rules licensing steps in a derivation on the basis of syntactic relations they bear to earlier steps.

In what sense is a model-theoretically defined logical consequence Q of P guaranteed to be true, if P is? Here we suppose that both are sentences

of a language of the standard first-order predicate calculus. It will then turn out (i) that it is impossible for the proposition expressed by P to be true without the proposition expressed by Q also being true, and (ii) that ⌜P → Q⌝ expresses a proposition that is knowable a priori to be true. These results tell us something positive about the strength of the guarantee we have that Q is true when it is a logical consequence of a true sentence P. Although the results about necessity and apriority don't *follow* from the model-theoretic definition of first-order logical consequence alone, they do indicate that the relation on sentences it defines has properties important for understanding the argumentative commitments we use it to track.[38] What does follow from the model-theoretic definition is that the sense in which the truth of P guarantees the truth of Q is independent of the special subject matter of P and Q. If we require Q to be true in every model in which P is true, then this condition is met, since the truth of P will be enough to determine the truth of Q in a (nonempty) domain of any size with nonlogical vocabulary assigned interpretations in the domain any way. Otherwise put, if Q is false in some models in which P is true, then one must appeal to special facts about the subject matter of P and Q (involving the domain of quantification or the denotations of nonlogical vocabulary) to get from the *actual truth* of P to the *actual truth* of Q—thus disqualifying Q from being a logical consequence of Q. For these reasons, the model-theoretic definition provides us with a notion of first-order logical consequence that is quite significant.

Whenever a formal system of derivability is presented, it is always relevant to ask whether all and only the proper derivations it defines are genuine instances of model-theoretically defined logical consequence. Using the definition, we can formulate such questions with mathematical precision. Just as one can mathematically investigate and answer questions like (21)–(23) about a theory of some particular subject matter, so one can mathematically investigate and answer questions like (24)–(26) concerning logical notions.

24. Is it the case that for all sentences P and Q, if Q is derivable from P in the formal system of proof S, then Q is true in all models in which P is true? (Is S sound?)
25. Is it the case that for all sentences P and Q, if Q is true in all models in which P is true, then Q is derivable from P in the formal system of proof S? (Is S complete?)
26. If it is not the case that for all sentences P and Q, Q is true in all models in which P is true iff Q is derivable from P in the formal system of proof S, is there any other system in which the equivalence does hold?

[38] Model-theoretically defined second-order logical consequence does not—or at any rate does not clearly—vindicate the idea that if Q is a logical consequence of P, then ⌜P → Q⌝ expresses a proposition that is knowable a priori to be true. See Soames (1999), pp. 130–36.

The fact that Tarski's definition of truth provided the foundation for productive investigations of metatheoretical questions like (21)–(26) is the main reason his conception of truth has been judged to be theoretically fertile. Thus, it is understandable why many have believed his conception of truth to be a good *explication* of our pretheoretic concept of sentential truth. Unfortunately, it isn't.

Although Tarski's definition of truth did provide a basis for the now standard definition of truth in a model, and the model-theoretic definitions of logical properties and relations, the sense in which it provided that basis belies its original status as a *definition* of truth. To see this, one must understand what the notion of *truth in a model* really amounts to. The technical concept defined is a two-place relation between sentences and models, which are pairs of a set D_M and a valuation function V_M. Although the notion defined is purely formal, it is the mathematically stripped-down version of a background idea that is not purely formal, but genuinely semantic.

When combined with the definition of *truth in a model,* D_M is taken to be the range of the quantifiers, which is the common element in the interpretations of '∃' and '∀', while V_M specifies the denotations, i.e., interpretations, of the nonlogical vocabulary. When used in this way, a model is the formal counterpart of an *interpretation* of the nonlogical vocabulary of a formal language. The logical vocabulary is treated as having fixed interpretations, which are the standard logical operations: negation for '~', conjunction for '&', etc. In this way, the interpretation of the language as a whole is divided into a part that is allowed to vary, which is represented by the model, and an invariant part reflected in the clauses of the definitions of *the denotation of a term (relative to an assignment) in a model* and *the truth of a formula (relative to an assignment) in a model.* In short, *truth in a model* is really a stand-in for *truth in an interpretation*—i.e., *truth of a sentence when it is understood in a certain way.*

Since the notion of an *interpretation,* in the sense we are using it here, is a technical one, *truth in an interpretation* is too. But the notion of sentential truth it employs is ordinary. To say that a sentence is true, in the ordinary sense, is to say that what it is used to say—the proposition it expresses—is true. So, to say that a sentence S is true in a given interpretation iff so-and-so is to say that when S is understood in the way specified by the interpretation, the proposition S expresses a truth iff so-and-so. It is only because we presuppose this ordinary notion of truth that the "interpretations" we abstract from models interpret sentences at all. In short, what is really being defined when we define *the truth of a sentence in an interpretation* is not a special formal sense of 'truth', but a special formal sense of 'interpretation'.

When we conceptualize things this way, the sentence to be interpreted is treated merely as a syntactic object to be put to use. The truth conditions it receives are its interpretation. The process generating them starts by taking the concept *denotation* to be a semantic primitive and specifying its

extension for the names, function signs, and predicates that occur in the sentences of the language being interpreted. The definition of *the denotation of a term in an interpretation relative to an assignment* takes the interpretations of names and function signs as input and specifies the denotations that become the interpretations of compound terms. Here again, we are contributing to the definition of *interpretation*, not *denotation*. The definition of *the truth of a formula in an interpretation relative to an assignment* takes the denotations (interpretations) of predicates and the denotations (interpretations) of terms as input and specifies the truth conditions of formulas relative to assignments. These are the interpretations of the formulas. The truth conditions of sentences, which are their interpretations, are derived from the interpretations of formulas.

The sense in which models/interpretations *interpret* sentences is, of course, very spare. For most purposes in the philosophy of language, the philosophy of mind, and theoretical linguistics, we need interpretations of (uses of) sentences that are much richer than mere statements of truth conditions expressed by material biconditionals. But what we are given here is enough for most logical and mathematical purposes to which model theory is put, including the standard model-theoretic definitions of logical consequence and related notions. Because the truth of a sentence P in a model is a proxy for the (ordinary) truth of what P is used to say, when P is interpreted in a certain way, the explanation given above of how truth-theoretically defined logical consequence allows us to systematize and study our argumentative commitments is preserved by the model-theoretic definition.

5.3. Truth, Meaning, and Tarski's Pseudo-Semantic Conception of Truth

Throughout my discussion of Tarski I have made use of the widely shared assumption that there is a conceptual connection between truth and meaning, in virtue of which understanding the meaning of a sentence typically involves knowing the conditions in which it is true, and knowing the conditions in which a sentence is true typically provides some information about its meaning. Different philosophers have held stronger or weaker versions of this view. The simplest version was articulated by Donald Davidson in his influential article "Truth and Meaning," published in 1967.

> (T) S IS T IFF P
>
> What we require of a theory of meaning for a language L is that *without appeal to any (further) semantic notions* it place enough restrictions on the predicate "is T" to entail all sentences got from schema T when 's' is replaced by a structural description [a transparent Tarskian name] of a sentence of L and 'p' by that sentence.
>
> Any two predicates satisfying this condition have the same extension, so if the metalanguage is rich enough, nothing stands in the way of putting what I

> am calling a theory of meaning into the form of an *explicit definition* of a predicate "is T." But whether explicitly defined or recursively [i.e., inductively] characterized, it is clear that the sentences to which the predicate "is T" applies will be just the true sentences of L, for the condition we have placed on satisfactory theories of meaning is, in essence, Tarski's Convention T that tests the adequacy of a formal semantical definition of truth.
>
> The path to this point has been tortuous, but the conclusion may be stated simply: *a theory of meaning* of a language L shows "how the meanings of sentences depend upon the meanings of words" if it contains a (recursive) *definition of truth-in L.* . . . I hope that what I am doing may be described in part as defending the philosophical importance of Tarski's semantic concept of truth. . . .
>
> There is no need to suppress, of course, the obvious connection between a *definition of truth of the kind Tarski has shown how to construct,* and the concept of meaning. It is this: the definition works by giving necessary and sufficient conditions for the truth of every sentence, and to give the truth conditions is a way of giving the meaning of a sentence. *To know the semantic concept of truth for a language is to know what it is for a sentence-–any sentence—to be true, and this amounts, in one good sense we can give to the phrase, to understanding the language.*[39]

If the view here expressed by Davidson were correct, then the notion of truth defined by Tarski could play the central role in a theory of meaning for the object language over which the predicate is defined. If such a result could be established, it would support the claims in Carnap (1942) and Tarski (1944) that Tarski's notion of truth can be used to define and study semantic notions such as meaning and synonymy, thereby providing further vindication for taking his definition to be an adequate explication of the notion of truth. But no such result can be established. On the contrary, the idea that anything remotely along these lines could be correct was one of the most widely shared errors in the history of the analytic tradition.

To see this, imagine that 'e' is a name of the earth, that 'R' is a predicate applying to all and only round things, that 'T in L' is a Tarskian truth predicate, and that (27) is an instance of schema T that is derivable in the metalanguage from an explicit Tarskian definition of *T in L.*

27. 'Re' is T in L iff the earth is round.

Since 'T in L' is the *definiendum* of the definition, it can be replaced, with no alteration of content, by the *definiens* (which, in accord with the demands of Tarski, is free of any semantic notions). Performing the replacement yields (28).

28. [There is a set T_L such that 'Re' is a member of T_L, and for all sentences s of L, s is a member of T_L iff (i) s = ⌜Pt⌝ for some one-place predicate P and

[39] Davidson (1967 [2001]), pp. 23–24, my emphasis.

term t, and there is an object o such that P *applies*$_T$ to o and o is *denoted*$_L$ by t; or (i) clauses for 2, 3, . . . n-place predicates (and terms); or (ii) S = . . . clauses for truth-functional connectives . . . ; or (iii) s = . . . clauses for quantifiers . . .] iff the earth is round.

Since 'Re' is a sentence consisting of a one-place predicate followed by a term, we can simplify (28) by dropping the extraneous clauses in (i), (ii), and (iii). This gives us (29).

29. [There is a set T_L such that 'Re' is a member of T_L, and for all sentences s of L such that s = ⌜Pt⌝ for some one-place predicate P and term t, s is a member of T_L iff there is an object o such that P *applies*$_T$ to o and o is *denoted*$_L$ by t] iff the earth is round.

Next, we replace 'denotes$_T$' and 'applies$_T$' with the *definiens* provided by an explicit Tarskian definition of each. This yields (30).

30. [There is a set T_L such that 'Re' is a member of T_L, and for all sentences s of L such that s = ⌜Pt⌝ for some one-place predicate P and term t, s is a member of T_L iff there is an object o such that (i) t = 'e' and o = the earth, or t = 'm' and o is Mars, or . . . (one disjunct for each name in L) . . . , and (ii) P = 'R' and o is round, or P = 'M' and o is massive, or . . . (one disjunct for each predicate of L) . . .] iff the earth is round.

Supposing we can recognize trivial identities and nonidentities of expressions of L, we can simplify (30) by eliminating the nonidentities. This gives us (31), which, in turn, is trivially equivalent to (32).

31. [There is a set T_L such that (i) 'Re' is a member of T_L, and (ii) 'Re' = 'Re' and 'e' = 'e' and 'R' = 'R' and there is an object o such that o = the earth and o is round] iff the earth is round.
32. There is an object o such that o = the earth and o is round iff the earth is round.

None of these biconditionals provides any information about the meaning of the object-language sentence 'Re'. One could know the facts they express without knowing the first thing about what the sentence does, or doesn't, mean. Suppose one didn't know that 'Re' means in L that the earth is round, and one was considering the hypothesis that it means that the earth is not round. Given (27)–(32) plus instances of the a priori schema that *If s means in L that P, then 'T in L' is a truth predicate for L only if s is T in L iff P*, one could conclude that either 'T in L' isn't a truth predicate for L (and T_L isn't the set of true sentences), or 'Re' *doesn't* mean in L that the earth is not round. But without knowing the meanings of the sentences of L in advance, one couldn't determine whether 'T in L' was a truth predicate, and without knowing that, one could determine *nothing* about the meaning of 'Re' from a statement of its "Tarski-truth conditions."

The key point is that instances of schema (33a), which contain our ordinary truth predicate, are obvious a priori truths, whereas instances of (33b), which contain a Tarskian truth predicate for L, are neither obvious nor knowable a priori.[40]

33a. If s means in L that P, then s is true in L iff P.
b. If s means in L that P, then s is T in L iff P.

It is the obviousness and availability of (33a) that allows claims of the form *s is true in L iff P* to provide information about meaning. If one knew that 'Re' is true in L iff the earth is round, then one could immediately eliminate the hypothesis that 'Re' means in L that the earth isn't round—since that hypothesis plus (33a) would contradict one's knowledge of the truth conditions of 'Re'. The unavailability of (33b) prevents similar conclusions from being drawn from claims of the forms *s is T in L iff P*. Consequently, those claims carry no information about meaning.

This result shows what should have been obvious all along: *Tarski's truth predicates aren't semantic*. The very fact that he required them to be definable entirely in terms of the non-semantic concepts expressed in the object language plus logic, set theory, and the syntax of L guaranteed that they couldn't be semantic. Semantic concepts are those intertwined with claims about meaning. Since no concepts definable from Tarski's antiseptically non-semantic base are semantic, the ubiquitous label applied to the notion he defined—*the semantic conception of truth*—is an absurd misnomer. It is a testament to the monumental historical misunderstanding of Tarski (1935) that the only major philosopher of his era who, to my knowledge, recognized this point was Alonzo Church.[41]

What is the source of the conceptual connection between truth and meaning that is missing in Tarski's substitute for truth? I believe it is the primacy of propositions as bearers of truth in the ordinary sense. The bearers of truth are, in the first instance, what is asserted and believed, which agents entertain and commit themselves to when they assertively utter, or otherwise accept, sentences. Sentences are true only derivatively, when the linguistic rules governing their use, which constitute the meaning of the sentence, determine a single proposition, which is, in fact, true. Consequently, when we are told that a sentence is true (in the ordinary sense), we are given information about its meaning and the proposition routinely expressed when it is used.

When a sentence contains no indexical or other semantically context-sensitive element, there is often a single proposition determined by its linguistic meaning that is reliably, though not invariably, a constituent of the illocutionary content of uses of the sentence. In these cases there is a close

[40] Instances are obtained by replacing 'P' with a sentence expressing a proposition.
[41] Church (1944), pp. 65–66.

relationship between talk about meaning of the sentence and talk about the proposition it expresses. In such cases, instances of schema (33a) are tantamount to instances of schema (34).

34. If S means in L—i.e., is used by speakers of L to express the proposition—that P, then the proposition expressed by S in L is true iff P.

The conceptual connection between truth and meaning is then the result of the fact that to say of S that it means in L that P is to say that uses of S express the proposition that P in L. This explains why to say that S is true in L is to say that the proposition expressed by S in L is true, and why instances of the propositional schema *the proposition that P is true iff P* are obvious, a priori, and necessary truths.

In sum, the information about meaning carried by statements specifying the truth conditions of sentences is due to the implicit commitment to propositions carried by ascriptions of our ordinary notion of truth to sentences. Roughly put, a sentence S of L is true iff uses of S in accord with the linguistic conventions governing S in L are true, where such uses either determine propositions or are themselves propositions. Since propositions play no role in the definition of Tarski's truth-substitute, predication of his concept to a sentence carries no information about the sentence's meaning. His predicate and the ordinary truth predicate of sentences do, of course, coincide in extension over the object language. But they don't express the same property, and so uses of sentences containing them don't carry remotely the same information.

The fact that Tarski's defined truth predicates are useless in semantics shows that his non-semantic notion of truth is not an adequate explication of our ordinary notion. But this doesn't mean that the recursive apparatus used in his characterization of truth, and truth in a model, is useless. Far from it. That apparatus is simply not the heart of a definition of truth. Rather, it, or a descendant of it, is an essential part of theories or definitions that *employ* the ordinary notion of truth for special purposes. In logic and model theory the Tarskian formal apparatus is incorporated in defining what it means for a model to be taken as a genuine interpretation of the sentences of a formal language. In empirical theories of meaning it is part of the systematic assignment of the conditions in which a sentence is true in the ordinary sense. These are magnificent contributions. They simply aren't contributions of the sort that they have often been taken to be.

6. CARNAP'S FLAWED TARSKIAN EPIPHANY

In sections 5 and 6 of chapter 7, I discussed the difficulties illustrated in Hempel (1935) and Reichenbach (1938) that led Carnap, Neurath, Hempel, and Reichenbach, along with other logical empiricists, either to identify the concept *truth* with the concept *being highly confirmed*, or to reject

the former in favor of the latter. Tarski (1935) immediately changed this for Carnap, and eventually for most of the others as well. Upon learning Tarski's views, Carnap become convinced of their broad philosophical importance and the need to communicate them to a general philosophical audience. To that end, he suggested to Tarski that he lecture on truth at the International Congress for Scientific Philosophy held in Paris in September of 1935. He reports Tarski as being skeptical that many philosophers would be interested, a skepticism that Carnap countered by promising to deliver his own lecture on the importance of Tarski's "semantic" conception.[42] So Tarski agreed to speak.

Carnap's recollection of the event is worth repeating.

> At the Congress it became clear from the reactions to the papers delivered by Tarski and myself that Tarski's skeptical predictions had been right. To my surprise, there was vehement opposition even on the part of our philosophical friends. Therefore we arranged an additional discussion. . . . There we had long and heated debates between Tarski, Mrs. Lutman-Kokoszynska, and myself on one side, and our opponents Neurath, Arne Ness, and others on the other. Neurath believed that the semantical concept of truth could not be reconciled with a strictly empiricist and anti-metaphysical point of view. Similar objections were raised in later publications by Felix Kaufmann and Reichenbach. I showed these objections were based on a misunderstanding of the semantical concept of truth, the failure to distinguish between this concept and concepts like certainty, knowledge of truth, complete verification and the like; I had already emphasized the necessity of this distinction in my Congress paper. Other misunderstandings and objections were clarified in a later article by Tarski [Tarski (1944)] and in my [Carnap (1946)].[43]

Carnap's Congress paper, Carnap (1935b), was translated into English and combined with Carnap (1946) to form Carnap (1949), the stated purpose of which was to clearly distinguish truth from confirmation. The paper begins on a promising note.

> The difference between the two concepts 'true' and 'confirmed' ('verified', 'scientifically accepted') is important and yet frequently not sufficiently recognized. 'True' in its customary meaning is a time-independent term. . . . For example, one cannot say "such and such a statement is true today (was true yesterday; will be true tomorrow)" but only "the statement is true." 'Confirmed', however, is time-dependent. When we say "such and such statement is confirmed to a high degree by observations" then we must add: "at such and such a time."[44]

Of course, a statement, i.e., a proposition, that is highly confirmed at one time may not be highly confirmed at another time, even though it is true

[42] See Carnap's "Intellectual Autobiography," in Schilpp (1963), p. 61.
[43] Ibid., p. 61.
[44] Carnap (1949), p. 119.

throughout. Hence, Carnap argued, the property *truth* of propositions is different from the property *being highly confirmed*. The main worry about the passage is that it doesn't identify statements with propositions. Problems will arise when Carnap removes this unclarity by identifying statements with sentences, which, by the inclusion of tense, may express different propositions with different truth values at different times. In such cases a sentence that would commonly be said to be true at one time would commonly be said to be false at another time. But that is a subject for later discussion.

However, the problem goes beyond sentences that are used to say different things at different times. The problem is Carnap's persistent conflation of sentences, uses of sentences in accord with the linguistic conventions of a language, and propositions. Being an opponent of conceptions of propositions as nonlinguistic entities that are the meanings of sentences, it was natural for him to use 'proposition' and 'statement' for *uses of sentences*. Although the idea that certain uses of sentences are propositions is perfectly justifiable, Carnap didn't systematically explore it, or inquire into what such uses are. Since it was also natural for him to take a sentence to be true iff *uses of it* are true, he was led to conflate sentences with propositions without paying attention to the difference between sentences as syntactic structures and cognitive uses those structures informed by linguistic conventions. This, as we shall see, spelled trouble. Whereas the definition of Tarski's truth predicate treated sentences as syntactic structures, abstracted from the semantic conventions governing them, the application of our ordinary notion of truth is parasitic on its application to uses of sentences in accord with their governing conventions.

Early in the article, Carnap recounts a bit of philosophical history leading up to 1935.

> [T]he concept of truth, when used without restrictions . . . leads to contradictions. . . . For this reason some logicians in recent times have been rather diffident in regard to this concept and have tried to avoid it. At times it was considered altogether impossible to establish an exact and consistent definition of truth (in its customary meaning); this has brought it about that the term 'true' was used in the entirely different sense of 'confirmed'. But this leads to considerable deviations from the common usage of language. Thus one would find it necessary to abandon, e.g., the principle of the excluded middle. . . . Tarski, however succeeded in establishing an unobjectionable definition of truth which explicated adequately the meaning of this word in common language [while adopting paradox-blocking restrictions]. Hence the term should probably no longer be used in the sense of 'confirmed'.[45]

Here Carnap correctly recants his earlier substitution of 'confirmed' for 'true', basing it, unfortunately, on the claim that Tarski's definition captures the ordinary meaning of 'true'.

[45] Ibid., pp. 119–20.

His reasoning is revealed by his comments about the following sentences (repeated here with his numbering):

[1] The substance in this vessel is alcohol.
[2] The sentence 'the substance in this vessel is alcohol' is true.
[3] X knows (at the present moment) that the substance in this vessel is alcohol.
[4] X knows that the sentence 'the substance in this vessel is alcohol' is true.

Carnap begins with a discussion of [1] and [2].

> Now the decisive point for our whole discussion is this: *the sentences [1] and [2] are logically equivalent*; in other words . . . they are merely different formulations of the same factual content; nobody may accept the one and reject the other; if used as communications, both sentences convey the same information though in different form. . . . It must be admitted that any statement of the logical equivalence of two sentences in English can only be made with certain qualifications, because of the ambiguity of ordinary words, here the word 'true'. The equivalence holds certainly if 'true' is understood in the sense of the semantical concept of truth. I believe with Tarski that this is also the sense in which the word 'true' is mostly used both in everyday life and in science.[46]

At this point Carnap delivers what he takes to be his argumentative clincher.

> The sentences [1] and [3] obviously do not say the same. This leads to the important result, which is rather obvious but often overlooked, that *the sentences [2] and [4] have different contents*. [3] and [4] are logically equivalent since [1] and [2] are. It follows that [2] and [4] have different contents. It is now clear that a certain terminological possibility cannot be accepted. "If we constantly bear in mind that the acceptance of any proposition may be reversed," in other words that [knowledge is never absolutely certain, but can have only a high degree of assurance capable of being weakened by future experience], "then we might instead call an accepted proposition a true proposition." This usage, however, would be quite misleading because it would blur the fundamental distinction between [2] and [3].[47]

Presumably the "terminological possibility" being rejected here is that a true proposition—for Carnap a true sentence—is a sentence used to make a statement that one has sufficient evidence to support a provisional claim to knowing. That is unacceptable, according to the passage, because it runs together [2], [3], and [4].

What is interesting about the passage is not the obviously correct conclusion Carnap endorses, but the theses about truth and Tarski-truth that he advances along the way.

[46] Ibid., pp. 120–21.
[47] Ibid., p. 121.

T1. Our ordinary predicate *true* of sentences of a language L means essentially the same thing as the predicate $true_{Tarski}$ Tarski defines.
T2. ⌜'P' is true⌝ and ⌜'P' is $true_{Tarski}$⌝ are logically equivalent to the sentence P, and so to each other. They are "different formulations of the same factual content"; "nobody may accept the one and reject the other"; and "they convey the same information."
T3. ⌜John knows that 'P' is true⌝ and ⌜John knows that 'P' is $true_{Tarski}$⌝ are logically equivalent to ⌜John believes that P⌝, and hence to each other.

To evaluate these theses, one must realize that both *true* and $true_{Tarski}$ can meaningfully be predicated of sentences one doesn't understand.[48] I can say, of a Japanese sentence, or a sentence of English that contains a word I don't understand, that I know, from the testimony of others, that it is true. I can do this using ⌜'P' is true⌝, which I understand perfectly well. By the same token, I can understand what ⌜'P' is $true_{Tarski}$⌝ means without understanding P.

The falsity of T1–T3 is now obvious. Suppose that ⌜'P' is $true_{Tarski}$ iff Mary gave Bill the book⌝ is a logical consequence of the Tarskian definition of $true_{Tarski}$ (as some such biconditional must be if the fragment of English in which the definition is constructed has, as Tarski requires, a paraphrase of each sentence of the object language). Then ⌜'P' is $true_{Tarski}$⌝ and 'Mary gave Bill the book' will be logical consequences of each other, and anyone who understands both will be in a position to logically derive one from the other.[49] By contrast, 'Mary gave Bill the book' is not a logical consequence of ⌜'P' is true⌝, nor is ⌜'P' is true⌝ a logical consequence of 'Mary gave Bill the book'. One could understand both and not be able to derive either one from the other. Hence, ⌜'P' is true⌝ and ⌜'P' is $true_{Tarski}$⌝ are logically independent; neither is a logical consequence of the other. This falsifies all three theses.

The seriousness of this error is illustrated by the following examples.

35a. John knows *that Mary gave Bill the book iff Mary gave Bill the book*.
b. John knows *that 'Mary gave Bill the book' is true iff Mary gave Bill the book*.

[48] 'P' is here used as a metalinguistic variable over sentences of the object language.

[49] Of course, if the definition of '$true_{Tarski}$' pairs ⌜'P' is $true_{Tarski}$⌝ with a sentence Q that is not identical with P but means the same as P, then, given the failure of the Transparency of Sameness and Difference of Meaning, there may be no guarantee that understanding both P and ⌜'P' is $true_{Tarski}$⌝ will allow one to see that accepting one calls for accepting the other. There may also be no guarantee that they will be logically equivalent, since sameness of meaning is (arguably but I think correctly) insufficient for logical equivalence. (See the discussion in section 3.4 of this chapter and think of the substitution of coreferential proper names or codesignative simple natural kind terms.) Thus, in addition to being wrong about the equivalence of ⌜'P' is true⌝ and P, and of ⌜'P' is $true_{Tarski}$⌝ and ⌜'P' is true⌝, Carnap was also not strictly correct in assuming the equivalence of ⌜'P' is $true_{Tarski}$⌝ and P. The correct point is that there will be some Q that means the same as P such that ⌜'P' is $true_{Tarski}$⌝ and Q are equivalent. Not so with truth, of course.

c. If John knows *that 'Mary gave Bill the book' is true iff Mary gave John the book*, then John knows enough to conclude *that 'Mary gave Bill the book' doesn't mean that Mary didn't give Bill the book*.

36a. John knows *that Mary gave Bill the book iff Mary gave Bill the book*.

b. John knows *that 'Mary gave Bill the book' is true$_{Tarski}$ iff Mary gave Bill the book*.

c. If John knows *that 'Mary gave Bill the book' is true$_{Tarski}$ iff Mary gave Bill the book*, then John knows enough to conclude *that 'Mary gave Bill the book' doesn't mean that Mary didn't give Bill the book*.

First consider (35). Obviously (35a) doesn't entail (35b). If it did, then the truth of (35a) would provide the same warrant for the truth of (35c) as the truth of (35b) does. That is absurd; (35b) warrants (35c), but (35a) doesn't. Knowledge of truth conditions provides one with nontrivial knowledge of meaning.

The examples in (36) illustrate why the same cannot be said about knowledge of Tarski-truth conditions. Since substitution of logical equivalents preserves logical equivalence, the complement clauses of (36a) and (36b) are, as Tarski and Carnap recognized, logically equivalent.[50] Since, according to Carnap in 1935, logical equivalents say the same thing, they can be substituted for one another in propositional attitude ascriptions. Hence, (36a) and (36b) are (for Carnap) equivalent, which is enough to show that one can know the Tarski-truth conditions of a sentence without knowing anything about its meaning. The point is underscored by the fact that for Carnap in 1935, (36c) is equivalent to the manifestly absurd (36d).

36d. If John knows that *Mary gave Bill the book iff Mary gave Bill the book*, then John knows enough to conclude *that 'Mary gave Bill the book' doesn't mean that Mary didn't give Bill the book*.

These examples show that statements of the Tarski-truth conditions play no role in endowing sentences with meaning, interpreting them, or describing their meanings once they have acquired them.

Carnap never saw this. After announcing in the preface of his book *Introduction to Semantics* that he was employing Tarski's notion of truth, he characterized the rules of a semantical system S (which are really substantive rules governing the use of its expressions) as constituting "nothing else than a definition of certain semantical concepts with respect to S, e.g., 'designation in S' or 'true in S'."[51] In section 7, on semantical systems, he says the following:

> A *semantical system* is a system of rules which state *truth-conditions* for the sentences of an object language and thereby determine the meaning of those sentences. A semantical system S may consist of *rules of formation*, defining

[50] This point is subject to the qualification explained in the previous footnote.
[51] Carnap (1942), pp. xii, 12.

> 'sentence in S', *rules of designation*, defining 'designation in S', and *rules of truth*, defining 'true in S'. The sentence in the metalanguage ⌜P is true in S⌝ means the same as the sentence P itself. This characteristic constitutes a condition for the *adequacy* of the definition.[52]

Note Carnap's insistence that the rules of the semantical system constitute *definitions* of 'designation in S' and 'true in S', exactly as Tarskian definitions define $denotation_{Tarski}$ and $true_{Tarski}$. Carnap adds that a metalanguage sentence that predicates truth of a sentence *means the same as* the sentence itself. This is false if by 'true in S' means *true in the ordinary sense*. It is true if (i) he means $true_{Tarski}$, (ii) what he calls "logical equivalence" is sufficient for sameness of meaning, and (iii) the metalanguage definition of $true_{Tarski}$ pairs P and ⌜'P' is $true_{Tarski}$⌝.[53] Carnap's final remark about the *adequacy* of the *definition* being provided by the equivalence of P and ⌜'P' is true⌝ leaves no room for doubt; by 'true' he means '$true_{Tarski}$'.

Carnap continues, describing semantic rules that

> determine a *truth-condition* for every sentence of the object language, i.e. a sufficient and necessary condition for its truth. In this way the sentences are *interpreted* by the rules, i.e. made understandable, because to understand a sentence, to know what is asserted by it, is the same as to know under what conditions it would be true. To formulate it in still another way: the rules determine the *meaning* or *sense* of the sentences.[54]

Here Carnap connects claims about truth conditions to claims about meaning and understanding. By contrast with the preceding paragraph just cited, this paragraph makes sense only if the truth conditions are stated using the ordinary truth predicate of sentences. If Tarski's defined truth predicate is intended, the remarks are absurd.

Two pages later Carnap is back with more claims that make sense only when the concepts at issue are *not* the genuine semantic concepts of designation and truth, but Tarski's non-semantic substitutes for them.

> By the rules of formation of a system S the term 'sentence of S' is defined; by the rules of designation 'designation in S'; by the rules of truth 'true in S'. The definition of 'true in S' is the real aim of the whole system; the other definitions serve as preparatory steps for this one, making its formulation simpler.[55]

A similar comment applies to the following passage.

> A remark may be added as to the way in which the term '*true*' is used in these discussions. . . . We use the term here in such a sense *that to assert that a sentence is true means the same as to assert the sentence itself*; e.g. the two statements "The

[52] Ibid., p. 22.
[53] The qualification in footnote 49 applies again here.
[54] Ibid., p. 22.
[55] Ibid., p. 24.

sentence 'the moon is round' is true" and "The moon is round" are merely two different formulations of the same assertion. (The two statements mean the same in a logical or semantical sense.)[56]

The appeal here to assertion, and its assumed connection with meaning, is interesting. In section 3.4 of this chapter I argued that, contrary to what one might first think, homophonic instances of *Schema True*—which connect English sentences S with ascriptions of the ordinary truth predicate to S—can't be known to be true or assertable simply by understanding them. I also argued that S and ⌜S is true in L⌝ neither mean the same thing, nor are necessary or a priori consequences of each other. But I didn't focus specifically on assertion, which is here leading Carnap astray.

Let S be the ordinary English sentence 'Five is a prime number'. Imagine it being used in an ordinary context in which speaker and hearer (i) understand the sentence, (ii) know that it expresses the proposition that five is a prime number and hence is true iff five is a prime number, (iii) presuppose this about each other, and (iv) realize that they both presuppose this. In this context, an agent who assertively utters *The sentence 'five is a prime number' is true* can correctly be reported either as having asserted *that 'five is a prime number' is true,* or as having asserted *that five is a prime number*—or as having asserted both. In many contexts, the same will hold for assertive utterances of 'Five is a prime number'. In these contexts it is transparent that to commit oneself to the truth of the sentence is to commit oneself to five's being a prime number, and conversely. Carnap was sensitive to this fact about assertive commitments, but he misdiagnosed its source. The sentences don't mean the same thing.

With this, I close the book on the error of assimilating ascriptions of our ordinary notion of truth, restricted to sentences of a given object language, to ascriptions of Tarski-truth to those sentences. To avoid this seductive error, one must avoid the following Carnapian mistakes:

(i) assuming, in setting up a "semantical system," a Tarskian definition of truth formulated in a fully meaningful metalanguage that already *contains* the object language, while at the same time treating the truth definition as endowing the object language with meaning;
(ii) assuming the truth definition will yield homophonic instances of *Schema T*, despite the fact that nothing in Tarski requires this, and inferring that, in general, instances of the schema are analytic, and so are necessary and a priori truths;
(iii) taking it for granted that homophonic instances of *Schema True*—which predicate the ordinary truth predicate to English sentences—can be known to be true and assertable simply by understanding them;

[56] Ibid., p. 26.

(iv) defining an analytic truth as one that can be known simply by understanding it, and concluding from (iii) that an ascription of the ordinary truth predicate to an English sentence S is analytically equivalent to S, and hence that the two are necessary and a priori consequences of each other.

(v) concluding from (ii) and (iv) that ascriptions of ordinary truth to S and ascriptions of Tarski-truth to S are conceptually equivalent to each other and to S.

It is important to note that my criticisms of the truth-conditional approach to meaning in Carnap (1942) target only one point—his misidentification of the notion of truth involved as Tarski's. Nothing has here been said against the general idea of truth-conditional theories of meaning that employ our ordinary notion of truth. Much of what Carnap has to say in *Introduction to Semantics* can be rendered coherent and interesting by substituting our ordinary notion of truth for Tarski's. Whether, in the end, versions of the truth-conditional approach to meaning in later work—by Carnap, Davidson, Montague, Kaplan, Lewis, Stalnaker, Chalmers, Jackson, and others—are successful will be assessed in later volumes. For now, it is sufficient to note that although the epiphany provided by Tarski's definition of truth liberated Carnap from his earlier non-semantic orthodoxy, and put him on a more productive path, it was based on a conceptual misunderstanding of what Tarski accomplished.

CHAPTER 10

Analyticity, Necessity, and A Priori Knowledge

1. LOGICAL EMPIRICISM'S LINGUISTIC CONCEPTION OF PHILOSOPHY AND THE MODALITIES

1.1. In What Sense Is Philosophy Supposed to Be Linguistic, but Not Empirical?

According to the *Tractatus*, every meaningful sentences falls into one of three classes—tautologies, that are true in virtue of meaning; contradictions, that are false in virtue of meaning; and synthetic sentences, the truth or falsity of which depend both on what they mean and on the way the world is. The logical empiricists made the same distinction, using the label 'analytic' for sentences that are true in virtue of meaning and 'empirical' for sentences the truth or falsity of which depend on both what they mean and the way the world is. They also identified analytic truths with

those that are necessary and knowable a priori (if they are knowable at all), while identifying empirical truths with those that are contingent and knowable only a posteriori.

In the *Tractatus*, empirical truths were the domain of science and analytic truths, the domain of logic and mathematics. Forthrightly recognizing that this left no room for philosophical truths, Wittgenstein paradoxically denied that there were any. Not recognizing the specious category of *important non-sense*, logical empiricists like Rudolf Carnap and A. J. Ayer didn't wish to characterize their voluminous output as any sort of nonsense. But avoiding that conclusion without repudiating central doctrines of logical empiricism wasn't easy.

Responses to this difficulty typically involved one or another, or some combination of three views. Sometimes logical empiricist doctrines were seen as analytic truths involving concepts and propositions of logical or scientific importance. Sometimes they were treated as empirical reports of the linguistic conventions governing the use of words and sentences in scientific and/or everyday contexts. Sometimes they were taken to be partly descriptive, partly normative *explications* of ordinary concepts like *meaning* designed to better fulfill the scientific purposes to which such concepts were put. The difficulties involved in giving a coherent answer to the question at issue will be illustrated by looking at what Ayer and Carnap had to say about it.

1.1.1. ACCORDING TO AYER

Ayer discusses his linguistic conception of philosophy in chapters 2 and 3 of *Language, Truth, and Logic* (1936). His general position is first stated as follows:

> [T]he propositions of philosophy are not factual, but linguistic in character—that is, they do not describe the behaviour of physical, or even mental objects; they express definitions or formal consequences of definitions.[1]

Though Ayer's general intent is clear, it's not easy to pin down the exact content of his remark. Although he speaks of propositions, he does not mean non-linguistic objects of the attitudes and bearers of truth value; he means sentences, or uses of sentences, without indicating which. I will call them *linguistic propositions*, leaving it open for now which is the better interpretation. In saying that philosophical propositions aren't *factual*, Ayer seems to be saying they aren't empirical, a posteriori truths (or falsehoods), which leads one to think they must be analytic truths (or falsehoods). That fits his description of them as expressing definitions or being consequences of definitions. Still, it is worrisome that Ayer explicates *being linguistic in character* as *not describing physical or mental objects*.

[1] Ayer (1936 [1946]), p. 57.

This is questionable. For one thing, even obviously analytic truths like *For all x, if x is an unmarried man, then x is not a married man* can be said to describe every object—as having a property that nothing could fail to have. Even if Ayer has in mind a special kind of description, merely ruling out descriptions of physical and mental objects leaves it open that philosophical statements might describe other kinds of objects, for example, linguistic objects. If philosophical statements can describe them, might the descriptions be empirical? One thinks he would say "No," but matters are not so simple.

Ayer follows up his general remark with the following statement:

> It follows that philosophy does not in any way compete with science. The difference in type between philosophical and scientific propositions is such that they cannot conceivably contradict one another. And this makes it clear that the possibility of philosophical analysis is independent of any empirical assumptions.[2]

Assuming that science can, in principle, pronounce on any contingent matter of fact, these remarks would seem to preclude any empirical statement—i.e., any contingent statement that is capable of being known to be true, or to be false, on the basis of evidence—from counting as a philosophical statement.

Although this comports with what seems to be Ayer's general intent, the next passage reveals a difficulty with this position that stems from his assumption that "propositions" are in some way linguistic.

> What has contributed . . . to the prevalent misunderstanding of the nature of philosophical analysis is the fact that propositions and questions which are really linguistic are often expressed in such a way that they appear to be factual. [Ayer here cites Carnap (1934b).] A striking instance of this is provided by the proposition that a material thing cannot be in two places at once. This looks like an empirical proposition, and is constantly invoked by those who desire to prove that it is possible for an empirical proposition to be logically certain.[3]

Throughout the book, Ayer takes a "proposition" to be *logically certain* iff it must, by its very nature, be true, and, hence, can be known to be so without appeal to empirical facts for justification. One wonders what sort of linguistic entity has those properties. Ayer continues:

> But a more critical inspection shows that it is not empirical at all, but only linguistic. It simply records the fact that, as a result of certain verbal conventions, the proposition that two-sense contents occur in the same visual or tactile sense field is incompatible with the proposition that they belong to the same material thing. And this is indeed a necessary fact.[4]

[2] Ibid., p. 57.
[3] Ibid., pp. 57–58.
[4] Ibid., p. 58.

The analysis of the proposition about material objects assumed in the passage comes from Ayer's problematic view that material objects are *logical constructions out of sense data*.[5] Fortunately, the content of that analysis is irrelevant to the points at issue here. Our concern is with his claim that certain facts about linguistic propositions are necessary.

The passage continues:

> But it has not the least tendency to show that we have certain [i.e., a priori] knowledge about the empirical properties of objects. For it [the proposition in question] is necessary only because we happen to use the relevant words in a particular way. There is no logical reason why we should not so alter our definitions that the sentence "A thing cannot be in two places at once" comes to express a self-contradiction instead of a necessary truth.[6]

A certain linguistic proposition—*that a material thing can't be in two places at once*—is here said to have the property of being necessarily true. But surely, one supposes, it has this property only if it has the property *being true* necessarily. Why then—since necessity and apriority go hand in hand for Ayer—shouldn't we conclude that here we have a priori knowledge that an object has an empirical property? Since truth is a property that some propositions have merely contingently, and which we can't know those propositions to have a priori, shouldn't it count as an empirical property? Now it so happens that Ayer is a redundancy theorist, and so could reply that truth isn't a property at all. Aside from being false, such a response is not at issue here, nor does Ayer invoke it. Instead, he points out that whether a sentence is true depends on what it means, which is determined by the linguistic conventions that govern it. Since what conventions do govern a sentence is a contingent matter, it isn't necessary that a sentence means what it does. Thus, it can't be necessary or a priori that any *sentence* is true.

These points, though correct, leave us with a mystery. What are the bearers of (necessary or contingent) truth, or falsity? This is a version of the problem raised in chapter 2 for Wittgenstein's conception of propositions in the *Tractatus*. The solution here is the same as it was there. We can approximate what both philosophers were after by taking propositions to be *uses of sentences in accord with certain conventions*. On this conception, it will be both necessary and knowable a priori that any use of a sentence in accord with certain conventions will be true.

But now there is trouble on another front. It will be contingent and knowable only a posteriori that any sentence we can identify has been used to state the necessary truth that a material object can't be in two places at the same time. This is significant because Ayer recognizes that identifying significant necessary a priori truths used in science and everyday life is a

[5] See Soames (2014), chapters 11 and 12 for an explanation and critique of this doctrine.
[6] Ayer (1936 [1946]), p. 58.

central task of philosophy. Although there is nothing inherently wrong with the idea that such identifications are empirical, it does threaten Ayer's general contention that philosophical propositions are never empirical.

Ayer illustrates his conception of philosophical analysis by using Russell's theory of descriptions as a paradigmatic example. He sees it as telling us

> every sentence which contains a symbolic expression of this form ["the so-and-so"] can be translated into a sentence which does not contain any such expression, but does contain a sub-sentence asserting that one, and only one, object possesses a certain property, or else that no one object possesses a certain property. Thus, the sentence "The round square cannot exist" is equivalent to "No one thing can be both square and round"; and the sentence "The author of *Waverley* was Scotch" is equivalent to "One person, and one person only, wrote *Waverley*, and that person was Scotch." . . . The effect of this definition of descriptive phrases, as of all good definitions, is to increase our understanding of certain sentences.
>
> And this is a benefit which the author of such a definition confers not only on others, but also on himself. . . . In general, we may say that it is the purpose of a philosophical definition to dispel those confusions which arise from our imperfect understanding of certain types of sentences of our language.[7]

One way of understanding these remarks is to see them are reporting contingent, a posteriori claims about the semantic properties of a certain class of English sentences that even sophisticated speakers of English might be confused about. Another way of taking some of them is to see them as elliptical for talk about the necessary and a priori consequences of certain uses of these sentences. It is, one might argue, a necessary, a priori truth that *any use* of 'The author of *Waverley* is a Scot' in accord with such-and-such conventions, and corresponding uses of 'Someone is both a Scot and identical with anyone who wrote *Waverley',* are necessary and a priori consequences of each other. Ayer is clearly more sympathetic to this second way of putting things. However, when he talks about the benefits of philosophical analyses like Russell's, he presupposes that all of us empirically identify such uses routinely, in his text, in our own speech, and in the speech of others. Is there any reason for banning explicit claims of this sort from philosophy because they are contingent and a posteriori?

Apparently, Ayer thinks so.

> It is misleading, also, to say, as some do, that philosophy tells us how certain symbols are actually used. For this suggests that the propositions of philosophy are factual propositions concerning the behavior of a certain group of people, and this is not the case. The philosopher who asserts that, in the English language, the sentence "The author of *Waverley* is Scotch" is equivalent

[7] Ibid., pp. 61–62.

> to "One person, and one person only, wrote *Waverley*, and that person was Scotch" is not asserting that all, or most, English-speaking people use these sentences interchangeably.[8]

In this passage Ayer eliminates a straw man. As we know from the extensive practice of today's theoretical linguists, the empirical claim that two natural-language sentences have the same semantic content is *never* identified with the claim that all, or most, of the speakers of the language use them interchangeably. Rather, the claim is related to the observations that verify a linguistic theory in the indirect way that is common to theoretical claims in empirical theories generally.

Next Ayer tries to identify the non-empirical claim that he thinks the philosopher does make.

> What he [the philosopher] is asserting is that, in virtue of certain rules of entailment which are characteristic of "correct" English, every sentence which is entailed by "The author of *Waverley* is Scotch," in conjunction with a given group of sentences, is entailed also by that group, in conjunction with "One person, and one person only, wrote *Waverley*, and that person was Scotch."[9]

Put aside whatever Ayer might mean by "certain rules of entailment." Let them be semantic conventions, whatever they are, that govern the use of sentences *in English*. He then identifies two English sentences and makes a claim about the logical relations holding between them and others. So understood, his statement is factual—i.e., contingent and knowable only a posteriori. It remains so, even if we recast it as about real, identified *uses of the two sentences*, in accord with conventions associated with them by speakers of English, by some group of English speakers, or even by Ayer himself. Supposing that Ayer could fully articulate all the relevant conventions, his statements would still be empirical, so long as actual uses of language were identified as being in accord with those conventions.

Although Ayer seems to recognize this, his recognition doesn't save his thesis that philosophical statements about language are never empirical.

> That English-speaking people should employ the verbal conventions that they do is, indeed, an empirical fact. But the deduction of relations of equivalence from the rules of entailment which characterize the English, or any other, language is a purely logical activity. And it is in this logical activity, and not in any empirical study of the linguistic habits of any group of people, that philosophical analysis consists.[10]

Here Ayer doesn't speak of propositions advanced as philosophical truths, but switches to speaking of philosophical analysis as an activity of deducing

[8] Ibid., pp. 69–70.
[9] Ibid., p. 70.
[10] Ibid., p. 70.

some things from other things. This is disappointing. For starters, it's not specific enough. Investigators in every field deduce many things from other things, without thereby doing philosophy. It also signals a retreat to the paradoxical tractarian position that because there are no philosophical truths, philosophy must be an activity aimed at something other than discovering truths. Although logical empiricists did sometimes say such things, it was widely supposed that the tractarian position could be avoided.

1.1.2. ACCORDING TO CARNAP

Interestingly, Ayer appends a footnote to the end of the passage just cited. It says:

> There is ground for saying that the philosopher is always concerned with an artificial language. For the conventions which we follow in our actual usage of words are not altogether systematic and precise.[11]

This remark is an accurate reflection of the practice of Carnap in *The Logical Syntax of Language* (1934b) (translated and expanded in 1937), *Introduction to Semantics* (1942), and later work—which we will sample in a moment. The point here is that appealing to an ideal language for which one can explicitly stipulate all relevant syntactic and semantic rules doesn't affect the issues raised by Ayer's discussion.

Suppose I stipulate rules for a version of the propositional calculus, thereby endowing the sentences with meanings and truth conditions. It will still be a contingent fact that they are governed by my stipulations and so have the meanings and truth conditions they do, which *you*, my audience, can know only a posteriori. Even I know them a posteriori—by knowing I have made the stipulations. What about stipulative utterances, e.g., *I stipulate that 'R' is to name Rudolf*? If I have the authority to stipulate, my statement can't fail to be true—not because it is analytic, but because to sincerely say one is stipulating that so-and-so is, within limits, to stipulate, and hence make it true, both that one is stipulating that so-and-so and that so-and-so. My knowledge of the resulting semantic properties of sentences is a posteriori, because it must be justified by my a posteriori knowledge of what I have done.

The upshot of this discussion is that much of what logical empiricists like Carnap called "*the logical analysis of the concepts and sentences of the sciences*," and identified with philosophy, can be seen as a special case of a descriptive empirical discipline we now call *theoretical semantics* plus the creative task of logico-linguistic innovation.[12] The task of philosophers like Carnap was in part to *describe* the logico-semantic properties of the expressions, formulas, and sentences of artificial languages already used

[11] Ibid., p. 70
[12] Carnap (1934a), and (1937), p. 292.

in logic, mathematics, and science, and in part to *add* to those languages in ways that would improve their utility in formalizing and systematizing discrete but related areas of science. Viewed in this way, it is striking how much the efforts of Carnap to advance his philosophical ends—especially his seminal work in formalizing the modalities—have now, and for many decades, been incorporated in highly theoretical *empirical* theories of the semantics of natural languages. This suggests that much of what he and other logical empiricists called *philosophy* was the embryonic stage of the young empirical science of linguistic semantics.

In the foreword of the original, 1934, German version of *The Logical Syntax of Language*, Carnap writes:

> That part of the work of philosophers which may be held to be scientific in nature—excluding the empirical questions that can be referred to empirical science—consists of logical analysis. The aim of logical syntax is to provide a system of concepts, a language, by the help of which the results of logical analysis will be exactly formulable. *Philosophy is to be replaced by the logic of science*—that is to say, by the logical analysis of the concepts and sentences of the sciences, for *the logic of science is nothing more than the logical syntax of the language of science*.[13]

These words, written in May of 1934, predated the (flawed) "semantic" epiphany that marked Carnap's response to Tarski's 1935 definition of truth. At this earlier time, he regarded logical concepts—which in later years he would attempt to define using Tarski's notion of truth—to be purely syntactic—i.e., proof-theoretic. The point of the book was to elucidate those concepts in a philosophy of logic that was pluralistic and conventionalist, incorporating a highly restricted logical core, while making room for different, noncompeting extensions resulting from the adoption of new linguistic conventions.

He summarizes this view in the following passage.

> [T]he view will be maintained that *we have in every respect complete liberty with regard to the forms of language; that both the forms of construction for sentences and the rules of transformation (the latter are usually designated as "postulates" and "rules of inference") may be chosen quite arbitrarily. . . . [T]his choice, whatever it may be, will determine what meaning is to be assigned to the fundamental logical symbols.* By this method, also, the conflict between the divergent points of view on the problem of the foundations of mathematics disappears. For language, in its mathematical form, can be constructed according to the preferences of any one of the points of view represented; so that no question of justification arises at all, but only the question of the syntactical consequences to which one or the other of the choices leads, including the question of non-contradiction. The standpoint which we have suggested—we will call it the Principle of

[13] Carnap (1934a), and (1937), p. xiii.

> Tolerance—relates not only to mathematics, but to all questions of logic. . . . In the domain of general syntax, for instance, it is possible to choose a certain form for the language of science as a whole, as well as for that of any branch of science, and to state exactly the characteristic differences between it and the other possible language forms.[14]

At this point, I am interested not in the specifically *syntactic* analyses that Carnap proposed in 1934, but in the view that philosophy, properly conceived, consists in the logical analyses (which may be truth-theoretic) of the concepts and sentences of science.

The key point, for our purposes, is that Carnap took genuine, philosophical statements—which he contrasted with pseudo-philosophical imposters—to be either explicit statements about the logical properties of linguistic expressions, or translatable, without loss of content, into such statements. Those that were not explicitly about expressions were said to be formulated in *the material mode of speech*. Though such statements were not denied legitimate uses, they were a source of serious error if they were taken to be what they superficially appeared to be—statements about objects other than linguistic expressions. Error was avoided by translating them into the *formal, i.e., explicitly linguistic, mode of speech*. When this could be done, use of the material mode was innocuous. When it couldn't be done, the statements were dismissed as pseudo-statements.

Nearly all of the fifth and concluding part of *The Logical Syntax of Language* is given over to translations designed to reveal the explicitly linguistic content of the study of *the logic science,* which was to be Carnap's replacement for philosophy. All of its statements are, for the reasons mentioned in discussing Ayer, contingent and knowable only a posteriori. The virtue in Carnap's position, as compared to Ayer's, is that Carnap took his enterprise to be one that "*takes the place of the inextricable tangle of problems which is known as philosophy*." In effect, it is a *science* the subject matter of which is the logical structure of the several sciences, and of science as a whole.

Thinking of it in this way does not require taking its claims to be either necessary or knowable a priori. For Carnap, all necessary, a priori truths are such because they are analytic sentences. Analyticity, in turn, is always the result of either explicit stipulation or implicit linguistic convention. Roughly put, a sentence (of a specific language) is analytic for Carnap iff it follows from the conventions or stipulations by which it is endowed with meaning. One would like to say that the claim one uses the sentence to make may be necessary and knowable a priori, even though the claim, *about the sentence*, that it is analytic, and hence necessary and a priori, is itself contingent and knowable only a posteriori, though Carnap himself didn't put it this way.

[14] Ibid., p. xv, my emphasis.

1.2. What Is It for a Sentence to Be True in Virtue of Meaning Alone?

Analytic sentences were said by logical empiricists to be true in virtue of meaning alone. Typically, this involved two claims. First, an analytic sentence was held to be one that could be known to be true, simply by understanding it. Since all of logic and mathematics was held to be analytic, it was not assumed that every analytic sentence is trivially recognizable by anyone who understands it, in the sense of being a competent speaker who understands its words and phrases plus the semantic import of its syntactic constructions. Nor, since analyticity was supposed to *explain* apriority, could one define analyticity as the property of being a sentence one could, in principle, come to know to be true by deducing a priori consequences of the information provided by one's understanding of the sentence.

Nevertheless, the general idea was something like that. According to Carnap, to understand a language was to know the conventions governing its expressions. These conventions involved tacit stipulations from which it followed, without appeal to further empirical information, that certain identifiable sentences are true, and certain rules of inference for generating sentences from other sentences are truth-preserving. On this picture, an analytic sentence is one the truth of which can be derived from the conventions one learns when one learns the language.[15] Because no use of the traditional philosophical notion of apriority is explicitly used in this characterization of analyticity, it was thought that the latter could, without circularity, be used to explain the former. This questionable idea will be scrutinized in section 3.

The second major thesis about analytic sentences was that their truth is entirely due to their meaning, in the sense that the state the world happens to be in makes no contribution whatsoever. By contrast, the truth value of any synthetic sentence is always the product of what it means—the way it represents the world as being—plus the way the world really is. It is true iff these coincide. The challenge was to find a way of understanding this contrast that would vindicate the idea that analyticity *explains* necessity. The idea may be illuminated using an analogy due to Gillian Russell.[16] If you know what it is to multiply by zero, then you know, when given zero plus another number to multiply, that it is irrelevant what the other number is. It's not that multiplication doesn't always require two arguments; it does. It's just that when one argument is zero, the other argument plays no role in the calculation. Similarly, one might argue, the truth value of a sentence is always a function of two arguments, its meaning and the state of the world. It is just that when the first argument is the meaning of an analytic sentence, the second argument is irrelevant. Any such sentence is

[15] See Carnap (1942), sections 14–16.
[16] Russell (2008).

true at all possible world-states, and so is a necessary truth, but the *reason* for this is that the world-states play no role in the calculation. The meaning of the analytic truth is sufficient by itself. Admittedly, one might still wonder whether all necessary truths are analytic, which logical empiricism also required. This will be an important question in volume 3. Here, we will turn to the reason that necessity and apriority were important to logical empiricists.

2. THE DOCTRINAL SIGNIFICANCE OF EXPLAINING NECESSITY AND APRIORITY LINGUISTICALLY

Although traditionally many philosophers have distinguished between analytic and synthetic statements, not all of them took the analytic/synthetic distinction to coincide with the necessary/contingent distinction and the a priori/a posteriori distinction. A necessary truth is a statement that is true, and could not have been otherwise. If a statement is necessary, then for any possible state w that the universe could be in, if the universe were (or had been) in state w, then the statement would be (or would have been) true. Following Wittgenstein, the logical empiricists held that all necessary truths are analytic, and that meaning was the source of necessity. The tractarian foundation of this view lay in Wittgenstein's idea that for a sentence to *say* anything, for it to provide any information, is for its truth to *exclude* certain possible states that the world could be in. Since necessary truths don't do that, they say nothing; and since they say nothing about the way the world is, the way the world is makes no contribution to their being true. Hence, it was thought, their truth must be due to their meaning alone.

As explained in the discussions of Carnap (1930/31, 1932, 1934) and Hahn (1933) in chapter 7, leading logical empiricists endorsed this tractarian reasoning. They also invoked a related idea. Being empiricists, they believed that all knowledge *about the world* requires empirical justification based on observation and sense experience. It follows that since a priori truths can be known without such justification, they must not be *about* the world. Thus, the logical empiricists reasoned, the world must play no role in determining that these statements are true. Rather, their truth must be due to their meanings alone. In short, the tractarian reasoning identified the necessary with the analytic, while the logical empiricist reasoning identified the a priori with the analytic.[17] In theory these could have amounted to different identifications, but in practice they didn't. There was no disagreement between Wittgenstein and the

[17] See chapter 4 of *Language, Truth, and Logic* (Ayer 1936 [1946]), which is fittingly titled "The *A Priori*."

logical empiricists on this point because both identified necessity with apriority. For these philosophers, the necessary, the a priori, and the analytic were one.

To this the logical empiricists added a claim of explanatory priority. The *reason* for the necessity or apriority of any statement is to be found in its analyticity. As they saw it, there is no explaining what necessity is, how we can know any truth to be necessary, or how our knowledge of any necessary truth can be a priori, without appeal to the notion of truth in virtue of meaning. Consider our knowledge that certain truths are necessary. The logical empiricists thought that without appeal to analyticity, one could make no sense of the idea of knowing that something not only is true, but would have been true no matter which possible state the world was in. Surely, they reasoned, we don't examine all possible world-states and evaluate the statement against them one by one. If, on the other hand, the truth of a statement is guaranteed by its meaning alone, then in knowing its meaning we know, or are in a position to come to know, that it must be true, no matter which state the world may be in. Hence, *knowledge of meaning* explains *knowledge of necessity*.

The logical empiricists made similar claims about a priori knowledge. According to the them, if p is necessary, then p is knowable a priori, and hence knowable independent of any possible confirmation or disconfirmation by experience. But how, these philosophers wondered, can any knowledge be independent of experience? Ayer raises this question at the beginning of chapter 4 of *Language, Truth, and Logic*.

> Having admitted that we are empiricists, we must now deal with the objection that is commonly brought against all forms of empiricism; the objection, namely, that it is impossible on empiricist principles to account for our knowledge of necessary truths. For, as Hume conclusively showed, no general proposition whose validity is subject to the test of actual experience can ever be logically certain.[18]

In calling a proposition/sentence *logically certain*, Ayer is here characterizing it as something which, by its very nature, can only be true, and, for that reason, can be known to be true without appeal to empirical facts for justification. In short, he is saying that no general proposition/sentence with empirical content is necessary and knowable a priori.

He continues,

> No matter how often it is verified in practice, there still remains the possibility that it will be confuted on some future occasion. The fact that a law has been substantiated in n – 1 cases affords no logical guarantee that it will be substantiated in the n^{th} case also, no matter how large we take n to be. And this means

[18] Ayer (1936 [1946]), p. 72.

> that no general proposition *referring to a matter of fact* can ever be shown to be necessarily and universally true. It can at best be a probable hypothesis.[19]

Ayer here contrasts being probable with being logically certain, which he identifies with being necessary and knowable a priori. When he speaks of a universal generalization that "refers to a matter of fact" he means, I think, one the truth of which depends on some *contingent* matter of fact. The reasoning is this: If a universal generalization makes a claim about the way the world actually is, then its truth depends on the contingent truth of all its instances. Since each of these can be known to be true only by experience, the future course of which we cannot know in advance, the generalization cannot be known with probability 1, and so, can neither be necessary nor knowable a priori.

He concludes:

> And this, we shall find, applies not only to general propositions, but to all propositions which have a *factual content*. They can none of them ever be logically certain [i.e., necessary and knowable a priori].[20]

Ayer's point is that if p is necessary, then it is knowable a priori, and hence has no factual content. The implication here is that if p has no factual content, then the world makes no contribution to its truth, in which case its truth must be due to its meaning alone.

This is made clear a few pages further on.

> There is no need to give further examples. Whatever instance we care to take, we shall always find that the situations in which a logical or mathematical principle might appear to be confuted are accounted for in such a way as to leave the principle unassailed. And this indicates that Mill was wrong in supposing that a situation could arise which would overthrow a mathematical truth.[21]

In short, Mill was wrong in denying that the propositions of mathematics are necessary, a priori truths.

> The principles of logic and mathematics are true universally *simply because we never allow them to be anything else*. And the reason for this is that *we cannot abandon them without contradicting ourselves, without sinning against the rules which govern the use of language*, and so making our utterances self-stultifying. In other words, *the truths of logic and mathematics are analytic propositions or tautologies*.[22]

According to Ayer, necessary truths are true no matter what state the world is in *because* they are true in virtue of meaning; similarly, they are knowable

[19] Ibid., p. 72, my emphasis.
[20] Ibid., p. 72, my emphasis.
[21] Ibid., p. 77.
[22] Ibid., p. 77, my emphasis.

a priori, without appeal to empirical evidence for justification, *because* this knowledge is knowledge of meaning. There is no philosophical mystery in our being able to know what we have decided our words are to mean. And surely, Ayer and other logical empiricists thought, there is no mystery in the idea that the truth of a sentence may follow, and be known by us to follow, entirely from our decisions about meaning. Putting these two ideas together, they thought that they had found a philosophical explanation of our a priori knowledge of necessary truths, which otherwise would have been problematic.

Although this picture was, for decades, attractive to many philosophers, it suffered from several problems that were not immediately apparent. First, it simply took for granted something that we now know requires argument—namely that all and only necessary truths are a priori truths. Second, the logical empiricist's claim that analyticity was conceptually prior to the notions of necessity and apriority, and could be used to give philosophically satisfying explanations of the latter, gave a hostage to fortune that would later be exploited. If the logical empiricist's conception of analyticity could be shown to be problematic—as Quine convincingly was to argue in "Two Dogmas of Empiricism" in 1951—then both necessity and apriority would be threatened, and the structure of logical empiricism would be undermined. Third, the logical empiricists seriously underestimated the conceptual difficulties in their own accounts of all the modalities, and, in particular, of the difficulties inherent in attempting to use knowledge of meaning to explain a priori knowledge. I will here concentrate on this third set of problems.

3. DID THE LOGICAL EMPIRICIST ACCOUNT OF THE MODALITIES REST ON A MISTAKE?

3.1. Underestimating the Differences between Sentences and Propositions

Analyticity—truth in virtue of meaning—is a property of *sentences*. If sentences express non-linguistic propositions, then those propositions can be analytic only in the derivative sense of being expressed by sentences that are. By contrast, when we speak of necessary truths as *statements* or *propositions* that are true, and would have been true no matter which possible world-state the universe was in, we cannot straightforwardly identify these statements or propositions with *sentences* used to express them. Since the meaning of a sentence is a contingent feature of it, there is no sentence that would have been true no matter which possible world-state the universe were in, because if the universe were in certain of those states, the sentence would mean something other than what it actually means. A similar point holds for a priori truths, thought of as those knowledge of which does not require justification by empirical evidence of any sort. Since knowledge of what sentences mean is never a priori in this sense, knowledge of their

truth is never a priori either. Thus, a priori knowledge *that so-and-so* can never be a priori knowledge that any sentence is true.

This is a *prima facie* problem for logical empiricism. Whether or not it can be overcome depends on how one understands claims that ascribe necessity and apriority to "statements." Carnap gives us some insight into his views on these matters in section V of *The Logical Syntax of Language,* where he distinguishes the material mode of speaking from the formal mode, endorsing the latter for logical and philosophical analysis. Since the claims *it is a necessary truth that . . .* and *it is knowable a priori that . . .* involve so-called "indirect discourse," we look to Carnap's analysis of such. It involves translating indirect discourse, which is a species of the material mode, into "direct discourse," which is a species of the formal mode. For example, he translates (1a) into (1c) via the intermediary of (1b) (which is also in the material mode).[23]

1a. Charles said (wrote, thought) that Peter was coming tomorrow.
b. Charles said a sentence which means that Peter is coming tomorrow.
c. Charles said the sentence 'Peter is coming tomorrow' (or a sentence of which this is a consequence).

About this, he says:

> The use of the indirect mode of speech is admittedly short and convenient, but it contains the same dangers as other sentences of the material mode. For instance, *sentence [1a], as contrasted with [1c], gives the false impression that it is concerned with Peter, while in reality it is only concerned with Charles and with the word 'Peter'*. When the direct mode of speech is used, this danger does not occur.[24]

The highlighted claim in this passage is remarkable. According to Carnap, (1a) gives the false impression of being about Charles and Peter, when in fact it is really about Charles and the name 'Peter'. Because Carnap takes this to be so, he thinks that (1c) captures what (1a) really means while avoiding the false suggestions to which (1a) gives rise. Nothing could be further from the truth. First, consider a counterfactual possibility in which Peter's comings and goings, and Charles's thoughts about them, are the same as they are at the actual world-state, but Peter—the one whose arrival is reported by our use of (1a)—is named 'Bill' and either no one is named 'Peter' or someone else is. Although the statement made by our actual use of (1a) would be true were that counterfactual possibility realized, the statement made by our actual use of (1c) would be false. Thus what is stated by our two uses is importantly different. Second, someone can know, of the statement made by our use of (1a),

[23] Carnap (1934a), and (1937), p. 292.
[24] Carnap (1937), p. 292, my emphasis.

that it is true, without knowing, of the statement made by our use of (1c), that it is true, and conversely. Thus, what Carnap has offered cannot be an analysis of (1a).

Although the intermediate translation target (1b) doesn't suffer from every problem with the putative analysis (1c) of (1a), it shares some of them. Suppose that on Wednesday Charles assertively uttered either "He is coming the day after tomorrow," or "He is coming on Friday," using 'he' to refer to Peter. Then (1a) will express a truth if uttered on Thursday. By contrast, (1b) will express a falsehood if uttered on Thursday because neither the *sentence* 'He is coming the day after tomorrow' nor the *sentence* 'He is coming on Friday' means that Peter is coming tomorrow.[25] What we need is something like *Charles used a sentence to assert that Peter is coming tomorrow*. But this takes us back to indirect discourse.

Next consider extending Carnap's "analysis" by translating reports like (2a) into reports like (2c).

2a. Charles knows that if Peter is coming tomorrow, then Peter is coming tomorrow.
b. Charles is warranted in accepting some sentence which means that if Peter is coming tomorrow, then Peter is coming tomorrow.
c. Charles is warranted in accepting the sentence 'if Peter is coming tomorrow, then Peter is coming tomorrow' (or a sentence of which this is a consequence)

Since this "analysis" shares the problems of the previous analysis, it can't be accepted. However, one who did accept it would naturally take knowledge of the meaning of the sentence '*if Peter is coming tomorrow, then Peter is coming tomorrow*'—i.e., knowledge of the Carnapian semantic conventions governing it—to warrant accepting that sentence. From here, it is a short step to the linguistic theory of the a priori. All that remains is to take *a priori knowledge* to be knowledge justified solely by virtue of understanding sentences, and to take sentences like (3c) to be "analyses" of sentences like (3a).

3a. Charles knows *a priori* that if Peter is coming tomorrow, then Peter is coming tomorrow.
b. Charles is warranted in accepting some sentence which means that if Peter is coming tomorrow, then Peter is coming tomorrow, simply by understanding its meaning.
c. Charles is warranted in accepting the sentence 'if Peter is coming tomorrow, then Peter is coming tomorrow' simply by understanding its meaning (or by understanding some sentence of which this is a consequence).

[25] Although uses of sentences containing indexicals express different propositions in different contexts of utterance, the linguistic meanings of the sentences don't change from one context of use to the next.

With this we see one line of reasoning that might have made the linguistic theory of the a priori appear plausible to Carnap and other logical empiricists. However, we also see that *if this, or anything like it, was the basis of the doctrine*, then the linguistic theory of the a priori rested on two mistakes—its faulty analysis of indirect discourse reports, and its replacement of the traditional conception of apriority as *that knowledge of which doesn't require justification by empirical evidence*, with *that which one is warranted in accepting merely by understanding it*.

3.2. An Alternate Route to the Linguistic Theory of the A Priori?

To get the linguistic theory of the a priori off the ground, without making the mistakes just indicated, one must recognize that when one says that *it is necessary, and knowable a priori, that all squares are rectangles,* what is said to be necessary and knowable a priori is not the sentence 'All squares are rectangles,' or any other. The challenge is to explain, how, in light of this, one is supposed to move from the claim that S is analytic to the claim ⌜it is necessary/knowable a priori that S⌝. In this section I will present an argument which, though unsuccessful, is not transparently absurd. Because of this, it, or something like it, may provide a partial explanation of why the doctrine seemed attractive to so many for so long. What we learn about the modalities by identifying the difficulties with the argument will also be important.

We begin by letting S be an analytic truth expressing proposition p.

(i) Since S is analytic, an agent can know that S expresses a truth by learning what it means.
(ii) The agent will thereby know the metalinguistic claim q—*that S expresses a truth*—on the basis of the evidence E provided by the agent's experience in learning the meaning of S.
(iii) Since the agent has come to understand S, the agent will also know, on the basis of E, that S expresses p (and only p).
(iv) Combining (ii) and (iii), the agent will thereby know, on the basis of E, that p is true. Since p follows from this claim, the agent will be in a position to come to know p.
(v) However, the claim that E *justifies*—by ruling out possibilities in which it is false—is not p, but q.
(vi) Since p can be known without justifying evidence ruling out possibilities in which it is false, there must be no such possibilities.
(vii) So, if S is analytic, p must be necessary, and (by the present reasoning) capable of being known to be so; p is also knowable a priori, since knowledge of p doesn't require evidence justifying p.

Though one might be fooled by this reasoning, if it were left implicit, the problems with it—apart from (i), which we here accept for the sake

of argument—can be clearly identified.[26] The most obvious difficulty concerns the knowledge of p reached at step (iv). Any agent who comes to know p by this route will know it *a posteriori*—that is, by appealing to the empirical evidence E used at steps (ii) and (iii) to justify the agent's conclusions. It is important to realize that in such a case the agent's actual knowledge of p will be *a posteriori* whether or not p is *knowable* a priori.

Worse, p will be knowable a priori only if there is a *different*, nonempirical, route to such knowledge. This undermines the point of the linguistic theory. For if there is another way of coming to know p, independent of one's knowledge of language or any other empirical truths, then the fact that p is expressed by an analytic sentence *plays no role in explaining the apriority of p*. With this in mind, one could afford to grant that an agent's knowledge of p could arise by the empirical route sketched in steps (i)–(iv). If it did, the agent would know p *a posteriori* even though p can also be known a priori. Even if the picture of knowing an a priori truth by this a posteriori linguistic route partially explains the appeal of the linguistic theory of the a priori, it does nothing to vindicate it.

This means that even if there are analytic sentences, in the sense in which the logical empiricists understood that notion, we still have no way of using such sentences to explain *any of our a priori knowledge*—let alone all of it. Now notice that the reasoning described in the argument by which an agent comes to know both p and the necessity of p requires the agent to employ a priori logical knowledge *independent of the linguistic conventions governing the sentence S* about which the agent is reasoning. So, even if there were no other problems with it, the argument would presuppose much of what the linguistic theory purports to explain.[27] This last point was the focus of the W.V.O. Quine's 1936 paper "Truth by Convention," which was to become, more than a decade after its publication, the historically most influential critique of the linguistic theory of the a priori. It is the subject of the next section.

4. OVERTHROW OF THE LINGUISTIC THEORY OF THE A PRIORI

The linguistic theory of the a priori rested on two bits of knowledge its proponents took to be unproblematic—(i) knowledge of what we have decided our words are to mean, and (ii) knowledge that the truth of certain sentences *follows from* our decisions about what the words they contain mean. However, there is a problem here, located in the words *follows from*. Clearly we don't stipulate the meanings of all the necessary/a

[26] See chapters 3 and 4 of Williamson (2008) for a catalog of well-taken worries about (i).
[27] In addition, (vi) falls afoul of the contingent a priori.

priori/analytic truths individually. Rather, it must be thought, we make some relatively small number of meaning stipulations, and then draw out the *consequences* of those stipulations for the truth of an indefinitely large class of sentences. What is meant here by *consequences*? Not wild guesses or arbitrary inferences, with no necessary connection to their premises. No, by *consequences* the logical empiricists meant *logical consequences*, *knowable a priori to be true if their premises are true*. But now we have gone in a circle. According to these philosophers, all a priori knowledge of necessary truths—including our a priori knowledge of the necessary truths of logic—arises from our linguistic knowledge of the basic conventions, or stipulations, that we have adopted to give meanings to our words. But to derive this a priori knowledge from our linguistic knowledge, one has to appeal to an antecedent knowledge of logic itself. Either this logical knowledge is a priori or it isn't. If it is a priori, then some a priori knowledge is not explained linguistically; if it is not a priori, then our knowledge of logic isn't a priori. Either way, the linguistic theory of the a priori fails.

That, in a nutshell, was one of the central arguments of Quine (1936). Although not fully appreciated when published, this argument eventually became a classic, and is now widely known for its powerful critique of the program of grounding a priori knowledge in knowledge of meaning. Since the problems with that program are even more severe than is sometimes realized, it may help to illustrate them with a simple example.

4a. For all x, if x is a square, then x is a rectangle with four equal sides.

Let us suppose that the word *square* means the same as the phrase *rectangle with four equal sides*. Then sentence (4a) is synonymous with, and expresses the same proposition as, (4b).

4b. For all x, if x is a rectangle with four equal sides, then x is a rectangle with four equal sides.

Next we distinguish two questions.

Q1. How do we know that (4a) is a true sentence of English?
Q2. How do we know that for all x, if x is a square, then x is a rectangle with four equal sides?

These are different questions. The knowledge Q2 asks about can be had by someone who knows nothing about the English language, whereas the knowledge that Q1 asks about is knowledge of a certain fact about English. Moreover, knowledge that (4a) is a true sentence of English is neither a priori, nor knowledge of a necessary truth. Rather, it is ordinary empirical knowledge of a contingent fact about our language—something one learns when one becomes a proficient speaker. By contrast, our knowledge that if something is a square, then it is a rectangle with four equal sides is a priori knowledge of a genuinely necessary truth.

Next we ask how, if at all, knowledge of meaning plays a role in answering Q1 and Q2. First consider Q1. If someone knows that *square* means the same as *rectangle with four equal sides*, then we may suppose that he or she knows that (4a) means the same as (4b), and hence that (4a) is true, if (4b) is. But how does such a person determine that (4b) is true? Well, it might be argued, (4b) is of the form *if p, then p*, and, surely, anyone who knows the meaning of *if, then* knows that any sentence of this form is true. But what exactly is it to know the meaning of *if, then*, and how is this knowledge used in determining that all sentences of the form *if p, then p* are true? Here, our attempt to use our knowledge of meaning to answer Q1 bottoms out in the question of how, if at all, our knowledge of the meanings of the logical operators explains our knowledge of which sentences are logically guaranteed to be true.

Next consider Q2. We may take it that our assumptions about meaning give the result that the proposition that for all x, if x is a square, then x is a rectangle with four equal sides is identical with the proposition that for all x, if x is a rectangle with four equal sides, then x is a rectangle with four equal sides. Since to know that so-and-so is just to bear the knowledge relation to the proposition that so-and-so, it follows that our knowledge that for all x, if x is a square, then x is a rectangle with four equal sides is simply our knowledge that for all x, if x is a rectangle with four equal sides, then x is a rectangle with four equal sides. So how do we know that? Well, it might be argued, to know that is just to know the proposition expressed by a logical truth of the form *if p, then p*, and, surely, anyone who knows the meaning of *if, then*, plus the meaning of the sentence replacing 'p', will know that proposition to be true. Again we may ask, what is it exactly to know this meaning, and how is this knowledge put to use to secure the desired result? Here, our attempt to use knowledge of meaning to answer Q2 bottoms out in the question of how, if at all, knowledge of meanings of the logical operators explains our knowledge of the propositions expressed by logically true sentences.

Faced with these questions, the standard move of the defender of the linguistic theory of the a priori was to claim (i) that logic is true by convention, and hence analytic, and (ii) that, therefore, knowledge of logical truth is nothing more than knowledge of meaning. (Similarly for knowledge that certain inferences are truth-preserving.) But these points are far from transparent, as can be seen by considering the following scenario. Suppose I were to introduce a simple logical language L by listing some predicates and names used in forming atomic sentences, plus the logical constants '&', 'v', '→', '~' and '∀', and the variables 'x', 'y', etc. Imagine that you already understand the names and predicates, but that the logical symbols are new to you. I next go on to endow the logical symbols with meaning by making a complicated stipulation of the following sort: Let these logical symbols of L mean whatever they have to mean to make true every sentence of each of the following forms

(A ∨ ~A), (A → A), [(A & B) → B], [A → A ∨ B)], [~(A & B) → (~A ∨ ~B)], [(A & (A → B)) → B],[∀x Fx → Fn], [∀x (Fx → Gx) & Fn) → Gn], etc.

The details of the stipulation are not important. The idea is to make a stipulation that can be satisfied only if '~' '&', '∀x' and all the other logical operators are assigned interpretations which assure that all and only those sentences of L that are standardly classified as logically true are guaranteed to be true by the meanings of the logical operators. Let us suppose, for the sake of argument, that this is possible. If some group or community decides to adopt such a stipulation as a linguistic convention governing their use of L, then it would be natural to characterize the logical truths of L as sentences that are *true by convention*, and hence, *analytic*.

So, at any rate, one might think; and, so far, we have found nothing to object to in that thought. But that isn't the end of the matter. What about (i) knowledge of which sentences of L are true by convention, and (ii) knowledge of the propositions expressed by those truths? Regarding (i), consider the sentence (4c) of L, which is a counterpart to the English (4b).

4c. ∀x (x is a rectangle with four equal sides → x is a rectangle with four equal sides)

To establish that this sentence is true by convention, one might reason as follows:

P1. All sentences of L of the form ∀x(*Ax* → *Ax*) are stipulated to be true, and so are true by convention.
P2. (4c) is a sentence of L of the form ∀x(*Ax* → *Ax*).
C. Therefore sentence (4c) is true by convention.

Similar arguments could be given for other logical truths of L.

Although there is nothing wrong with these arguments, each presupposes a certain logical fact. Each argument is of the form:

P1. All F's are G (All sentences of such-and-such a form are true).
P2. n is an F (n is a sentence of such-and-such a form).
C. Therefore, n is G (Sentence n is true).

In order for someone to recognize that the premises of the argument justify the conclusion that a certain sentence of L is true, that person must recognize that if all F's are G's, and n is an F, then n is a G.[28] *This knowledge* isn't explained by knowledge of any stipulations about L; rather it is presupposed in using knowledge of the stipulations to arrive at knowledge

[28] The point here is not, of course, that in order to draw the conclusion he needs the claim that if all F's are G's and n is an F, then n is a G as a further premise. (We know from Lewis Carroll that this isn't so.) The point is (i) that if he is to *know* the conclusion on the basis of knowing the premises, he must recognize the argument as *justifying* the conclusion, and (ii) that recognizing this counts as knowing that if all F's are G's and if n is an F, then n is a G.

of which sentences of L are true. Consequently, although (4c) can be regarded as a sentence of L that is true by convention, and although one can arrive at the knowledge that it is true by learning the linguistic conventions of L, one can do so only if one has *prior* knowledge of the truth of propositions expressed by logical truths of the form *if all F's are G's, and n is an F, then n is a G*. This is precisely the kind of genuine, a priori knowledge of necessary truths for which the logical empiricists promised an explanation. What we have seen is that in appealing to the linguistic conventions of L, they haven't succeeded in giving one.

The same point could be made by focusing on sentences of English that are logical truths, and the propositions they express. The only difference is that it now becomes even harder for defenders of the linguistic theory of the a priori to make their case. When introducing logical constants into the new language L by stipulation, I was free to express the stipulation using antecedently understood expressions of English, including logical terms like *every*. However, if we try to imagine all the logical terms in English getting their meanings by stipulation, we are at a loss to understand how such stipulations could be expressed. Thus, it is harder to understand in what sense the logical truths of English could be true by convention in the first place.

Perhaps this last difficulty isn't insuperable. Perhaps speakers have some beliefs and intentions independent of any ability to express them in language. Perhaps some of these language-independent beliefs and intentions are about the use of expressions, and the meanings that speakers intend to assign to them. If so, then a case might be made for holding that these beliefs and intentions have the effect of meaning-giving stipulations, even though they are not publicly expressed in language. If so, then someone might argue that the logical words, for example, acquire their meanings by such real but unexpressed stipulations, in which case it might be maintained that the logical truths of English and other natural languages are true by convention in some extended sense.

But even if this were so, speakers' knowledge that certain sentences are true (or true by stipulation) would still presuppose antecedent, a priori knowledge of logical facts—i.e., a priori knowledge of certain (necessarily true) propositions expressed by logically true sentences. Since this is precisely the sort of knowledge that the proponents of the linguistic theory of the a priori were trying to explain, it is difficult to see how they could succeed. Putting this in terms of answers to our illustrative questions Q1 and Q2, we see that although these philosophers were right in thinking that knowledge of meaning may play a role in answering Q1, they did not succeed in showing that such knowledge is sufficient by itself (without appeal to prior knowledge of logical facts) to answer Q1; nor were they able to show that it makes *any* contribution to answering Q2.

For all these reasons, the program of explaining a priori knowledge by appeal to analyticity and linguistic conventions did not succeed. Despite

Quine's arguments in "Truth by Convention," this was not widely recognized until he revisited the topic of analyticity many years later in "Two Dogmas of Empiricism," published in 1951.[29] By that time, crippling difficulties with the empiricist criterion of meaning had made it obvious that there were intractable difficulties at the center of the philosophical vision of the logical empiricists.

[29] Who was the primary target of the critique in Quine (1936)? Until recently, Carnap was the consensus choice. This is contested in Ebbs (2011), which maintains that Carnap wasn't a target. Although I don't think that's right, I have been persuaded by more recent scholarship that Quine's proximate target was a view expressed by his Harvard colleague C. I. Lewis in *Mind and the World Order* (1929). That said, there is still reason to think that Quine rightly took his critique to apply to Carnap as well.

CHAPTER 11

The Rise and Fall of the Empiricist Criterion of Meaning

1. THE PHILOSOPHICAL SIGNIFICANCE OF THE EMPIRICIST CRITERION OF MEANING

In the last chapter, I examined the idea that analytic sentences express necessary truths that are knowable a priori simply by understanding and reflecting on the meanings of the sentences that express them. A sentence was regarded as contradictory if and only if its negation was analytic. All other meaningful sentences were classified as synthetic, contingent, and knowable only empirically. The empiricist criterion of meaning focused on this last class of sentences.

Its guiding idea may be put as follows:

THE BASIS OF VERIFICATIONISM

A nonanalytic, noncontradictory sentence S is meaningful iff S bears relation R to sentences the truth or falsity of uses of which can be determined by simple observation.

The most important task facing the logical empiricists was to precisely define the relation R in this principle. At the outset, leading positivists, including Carnap, Schlick, and Ayer, underestimated how difficult

this would turn out to be.[1] They were confident they had discovered a fundamental insight that would transform philosophy, and, for the first time, put it on a solid foundation. The chief cause of past philosophical confusion, and the reason for the lack of more significant progress, was, in their minds, that previous philosophers hadn't realized that all meaningful sentences have to be either analytic, contradictory, or empirically verifiable. For that reason, many of their works, particularly in ethics and metaphysics, were filled with sentences that don't fall into these categories.

Metaphysical sentences aren't analytic, because the truth or falsity of uses of them is supposed to depend on more than their meanings. Since these uses purport to be about the world, their truth or falsity must be determined by whether or not they correctly describe it. Despite this, these statements were often held to be necessary and knowable independently of experience, in the sense that ordinary observation wasn't needed to ascertain their truth. The logical empiricists believed this combination of characteristics to be impossible. Any genuine claim that purports to be about the world must be both contingent and capable of being verified or falsified by experience. Since uses of sentences to make metaphysical statements don't pass this test, such sentences were rejected as meaningless. Their negations were also rejected. Thus, in proclaiming that 'God exists' is cognitively meaningless, logical empiricists didn't take themselves to be committed to saying "God doesn't exist." On the contrary, they maintained that if 'God exists' is meaningless, then 'God doesn't exist' is too. According to logical empiricism, there are no genuine metaphysical problems for metaphysical statements to address.

Similar points were made about ethical theories. Often, the most fundamental claims made by uses of ethical sentences had been regarded as necessary (and knowable a priori), if true at all. But the sentences used to make those claims hadn't been thought to be analytic, because accepting them involved more than deciding how to use words. Uses of them played important roles in guiding action, even though they were taken to be descriptive, and so capable of being true or false. Logical empiricists insisted this combination of properties was incoherent. For them, necessity and apriority sprang from analyticity, and no statement could be both a fact-stating description and an action-guiding admonition. Most took ethical sentences to be cognitively meaningless, and so incapable of being used to make statements or express genuine beliefs. At best they

[1] See above, (i) the discussion of the closing sections of Carnap (1928 [1967]) in chapter 6, section 6, and chapter 7, section 1, (ii) the discussion of Schlick (1930/31 [1959]) and Carnap (1932 [1959]) in chapter 7, section 2, and (iii) the discussion of Schlick (1932/33) and Schlick (1934 [1959]) in chapter 7, section 4. Ayer's early understanding of the empiricist criterion of meaning will be discussed below.

were seen as disguised imperatives used to make recommendations, or to give orders.

The fact that logical empiricists rejected entire domains of traditional philosophical inquiry didn't mean that they thought that all traditional philosophy was mistaken. They viewed some as having resulted in important linguistic clarifications. Hume's analysis of causation as constant conjunction, Locke's conception of all knowledge as arising from experience, Russell's theory of descriptions, his reduction of arithmetic to logic, and his theory of logical constructions, were viewed as milestones. Wittgenstein's attempt to trace the limits of the meaningful was seen as a breakthrough. But it wasn't viewed as a triumph of traditional philosophy. Rather, it was adopted as the chief document ushering in the new beginning in philosophy announced in the 1929 logical empiricist manifesto, "The Scientific Conception of the World," and trumpeted by Schlick, Carnap, and Hahn in early issues of *Erkenntnis*.

No principle was more important for what was to be the new era of scientific philosophy than the empiricist criterion of meaning. The first attempts to formulate it were based on the idea that an empirical—i.e., nonanalytic, noncontradictory—sentence is meaningful if and only if the truth, or falsity, the statement it is used to make could, in principle, be conclusively established by deriving it from true observation statements. Testing this idea involved (i) distinguishing sentences used to make observation statements from other sentences used to make empirical statements, and (ii) specifying the logical relationship between an empirical sentence S and a set O of observation sentences needed in order for uses of the sentences in O to verify, or to falsify, a use of S.

2. OBSERVATION STATEMENTS

The first step in trying to turn the informal idea behind verificationism into a precise criterion of meaning was to characterize the class of observation statements. As indicated in chapter 7, this was a bone of contention from the beginning, with different logical empiricists offering different characterizations at different times. The central dispute was over whether observation statements should be taken to be statements about one's own sense data (that one could not possibly be mistaken about), or whether ordinary (fallible) statements about perceivable, medium-sized, physical objects should count as observational. Schlick (1934) advocated the former position, Neurath (1932/33) the latter, and Carnap moved from being friendly to the former in Carnap (1928) to being friendly to the latter in Carnap (1932/33b).

Ayer was also originally attracted to the first, and more radically empiricist, alternative. In *Language, Truth, and Logic*, he takes sense data to be

objects of perception, and compounds Russell's error in *Our Knowledge of the External World* by declaring not only (i) that material objects are logical constructions out of sense data,[2] but also (ii) that other people are logical constructions out of material objects (i.e., statements about other minds are analyzable into statements about the behavior of other bodies).[3] Thus, he saddled himself with the view (iii) that both material objects and other people are logical constructions out of sense data. Whose sense data? Although Ayer doesn't deal with the question in much detail, the need to avoid circularity invites the thought that material objects and other people are logical constructions out of *one's own* sense data. The resulting doctrine then maintains that any statement one makes (and any thought one entertains) that might seem to be about material objects and other people is, really, a statement (thought) about one's own sense data, *and nothing more*—i.e., about sense data one is experiencing, has experienced, or would experience if various (solipsistically characterized) conditions were fulfilled. Avoiding this *reductio ad absurdum* forced Ayer, as it had forced Carnap in the *Aufbau,* to a starting point consisting of free-floating, agentless experiences, out of which both agents and the world are "constructed." The critique of Carnap's position in section 5.3 of chapter 6 applies with equal force to Ayer.

The way out of this dead end is to give up the view that material objects are logical constructions out of sense data. But if material objects are regarded as distinct from sense data, with only statements about the latter being regarded as observational, then verificationists will have trouble with material-object statements from the start. Since the material-object sentences used to make these statements are not logically entailed by any finite set of sentences used to make sense-data statements, they won't count as conclusively verifiable, and the empiricist criterion of meaning formulated in terms of conclusive verifiability or falsifiability will be threatened before it gets off the ground. Spirited disputes over these issues occupied the logical empiricists through the middle 1930s.[4] Eventually, however, the disputes faded in significance, as it became more widely accepted that sentences used to make statements about the observational properties of

[2] Ayer (1936 [1946]), pp. 63–68.

[3] Ibid., pp.128–32. On p. 130 Ayer says: "[T]he distinction between a conscious man and an unconscious machine resolves itself into a distinction between different types of perceptible behavior. The only ground I can have for asserting that an object which appears to be a conscious being is not really a conscious being, but only a dummy or a machine, is that it fails to satisfy one of the empirical tests by which the presence or absence of consciousness is determined. If I know that an object behaves in every way as a conscious being must, by definition, behave, then I know that it is really conscious. . . . For when I assert that an object is conscious I am asserting no more than that it would, in response to any conceivable test, exhibit the empirical manifestations of consciousness."

[4] See Neurath (1932/33), Carnap (1932/33b), Schlick (1934); Ayer (1936/37). See also section III of the introduction to Ayer (1959).

physical objects could play the role of the protocol sentences in terms of which verifiability and falsifiability were defined. A little later, when severe problems inherent in attempts to formulate the empiricist criterion of meaning were recognized, it became apparent that difficulties in defining the relationship that sentences used to make non-observation statements were supposed to bear to sentences used to make observation statements in order to count as empirically meaningful would remain, no matter how the original disputes over protocol sentences were resolved. Thus, for us, a liberal and informal characterization of observation statements will be sufficient.[5]

OBSERVATION STATEMENTS

An observation statement is one that could be used to record the result of a possible observation. These statements assert that specifically mentioned observable objects have, or lack, specified observable characteristics—e.g., *The book is on the table, The chalkboard isn't green, The cup is empty and the glass is full.*

I leave aside such questions as *Observable by whom?* and *Observable by what means?* Instances of ordinary, unaided observation by normal human beings count as possible observations that may be recorded in observation statements. Whether or not observations involving magnifying glasses, binoculars, telescopes, microscopes, radio telescopes, electron microscopes, etc., should be counted as observations for these purposes is a vexed issue. On one hand, logical empiricists didn't want to include among the observational any statements the verification of which required both sense experience and substantial theoretical assumptions to interpret that experience. On the other hand, it was up for grabs what should count as substantial theoretical assumptions. It was also up for grabs whether there is a single, principled way of drawing the distinction between observation and theory, or whether, instead, there are different, context-sensitive, ways of drawing the line in different situations, for different scientific or philosophical purposes.

These potentially important questions would have to be addressed, if we could construct otherwise unproblematic versions of the empiricist criterion of meaning. As it turns out, formidable obstacles prevent us from doing that, no matter how observation statements are defined. For this reason, I will proceed as if there were a principled distinction between observational and non-observational claims, without worrying too much about how or where, precisely, the line is to be drawn.

[5] This definition allows sentences of different logical forms to count as observational—e.g., simple atomic sentences, negations, conjunctions, and even (in special cases) universal generalizations. In what follows, when I contrast observation sentences with, e.g., universal generalizations, the contrast will be between observation sentences and universal generalizations that are not themselves observational.

3. EMPIRICAL MEANINGFULNESS AS CONCLUSIVE VERIFIABILITY OR FALSIFIABILITY

Conclusive verifiability and falsifiability were typically defined as follows.

CONCLUSIVE VERIFIABILITY: FIRST PASS

A statement S is conclusively verifiable iff there is some finite, consistent set O of observation statements such that O logically entails S.

CONCLUSIVE FALSIFIABILITY: FIRST PASS

A statement S is conclusively falsifiable iff there is some finite, consistent set O of observation statements such that O logically entails the negation of S.

The first point to notice about these definitions is that conclusively verifiable statements are not invariably true, and conclusively falsifiable statements are not invariably false. The desired result is that a statement is conclusively verifiable if and only if, in some possible circumstances, it could conclusively be shown to be true by virtue of the fact that it follows logically from a set O of observation statements. A similar point holds for conclusive falsifiability. A statement is conclusively falsifiable if and only if, in some possible circumstances, it could conclusively be shown to be false by virtue of the fact that its negation follows logically from a set of observation statements that could be jointly true. The requirement that O be consistent is meant to ensure that it is possible for its members to be jointly true. The requirement that O be finite is meant to guarantee that it is possible for us to perform the observations involved.

In putting things this way, I followed the example of the logical empiricists in equating logical consistency with an ordinary notion of possibility, despite the fact that the identification of the two is now widely, if not universally, rejected. One difference concerns the *bearers* of the two notions of possibility. The bearers of logical possibility, i.e., logical consistency, are, in post-Tarskian logic, *sentences* of formal languages. The bearers of epistemic or metaphysical possibility are statements made or propositions expressed when agents use sentences. This leads to variations in what is "possible" in the two senses. For example, an identity *sentence* in which two different proper names flank the identity sign is, along with its negation, *always* logically possible and *never* logically necessary. But the proposition expressed by using an identity sentence involving coreferential names is *always* metaphysically necessary and its negation is *always* metaphysically impossible. Similarly, the proposition expressed by a use of an identity sentence involving names referring to different things is *always* metaphysically impossible, and its negation is *always* metaphysically necessary. Ignoring these differences invites confusion.

Such confusion often arose from the tendency of logical empiricists to slip back and forth between claims about *sentences* and claims about *statements*, while being vague and elusive about the relationship between the

two. On the one hand, they didn't want simply to identify sentences and statements. After all, two people can make the same statement by uttering different sentences; moreover, in certain cases, the same sentence can be used to make different statements. On the other hand, the logical empiricists didn't want to say that statements are distinct from the sentences used to make them. On the whole, they were content to observe that just as statements are made using sentences, so the statement itself, that which is stated, is nothing more than a sentence *used in a certain way*. Though arguably on the right track, this way of putting the idea is problematic, in part because it is doubtful that there is any entity a-sentence-used-in-a-certain-way distinct from the sentence itself. There are, however, *uses of sentences*—e.g., acts of using this or that sentence, in accord with the linguistic conventions that govern it, to predicate a given property P of a given object o—that may count as statements made or propositions believed. As explained in chapter 2, there are also statements/propositions—predicating P of o, with or without a linguistic intermediary—abstractable from these.

We can use this distinction between sentences and uses of sentences to clarify the logical empiricists' attempt to formulate an acceptable criterion of meaning. It is *sentences* that are contingently meaningful or meaningless; their meanings are the conventions that govern their use. It is *uses of sentences* in accord with the conventions governing the sentences (and further statements/propositions abstractable from them) that are true or false, and that have their truth conditions essentially (rather than contingently). With this in mind, we can restate the original definitions of conclusive verifiability and conclusive falsifiability so as to avoid confusing sentences and statements.

CONCLUSIVE VERIFIABILITY S

A *use* of a *sentence* S in conformity with the linguistic conventions that govern S is conclusively verifiable iff there is some finite, consistent set O of *sentences* which, *when used in conformity with the linguistic conventions that govern them*, predicate observational properties of things, such that O *entails* S.

CONCLUSIVE FALSIFIABILITY S

A *use* of a *sentence* S is conclusively falsifiable iff there is some finite, consistent set O of sentences which, *when used in conformity with the linguistic conventions that govern them*, predicate observational properties of things, such that O *entails* the negation of S.

These reformulations avoid the incoherence of identifying the bearers of meaning and logical properties and relations with the true or false statements they are used to make. But that change does not, by itself, bring logical necessity and possibility into line with more ordinary notions of necessity and possibility. Doing that requires replacing the standard notion of *logical entailment* in the original definitions with a notion of *semantic entailment* that

incorporates meaning postulates in the sense of Carnap (1952). This is why I have replaced *logically entails* with *entails* in the new definitions.

A meaning postulate is a sentence of the object language the role of which is to capture conceptual, but nonlogical, relationships determined by the meanings of nonlogical vocabulary. Semantic entailment is model-theoretic entailment restricted to models (interpretations) that satisfy (make true) all meaning postulates. For example, some meaning postulates require pairs of synonymous predicates to be assigned the same extensions in every admissible model, some require pairs of coreferential proper names, and pairs of codesignative natural kind terms, to be assigned the same referents in every admissible model, and some require pairs of predicates that exhibit meaning inclusion to be assigned extensions one of which is a subset of the other in all admissible models. By these means, the definitions of conclusive verifiability and conclusive falsifiability can be made to generate results more closely aligned with our ordinary notions of possibility. Fortunately, the conclusions drawn about the proposals that follow are general enough to apply no matter what reasonable choices about meaning postulates are made.

We are now ready to consider two attempts to base empirical meaning—the meaning of nonanalytic and noncontradictory sentences—on conclusive verifiability and conclusive falsifiability.

ATTEMPT 1

A nonanalytic, noncontradictory sentence S is empirically meaningful iff uses of S, in conformity with the linguistic conventions that govern it, are conclusively verifiable.

ATTEMPT 2

A nonanalytic, noncontradictory sentence S is empirically meaningful iff uses of S, in conformity with the linguistic conventions that govern it, are conclusively falsifiable.

These two attempts come to grief over the following facts.

Fact 1: Uses of universal generalizations (and of negations of existential generalizations) are not conclusively verifiable.

(i) All moving bodies not acted upon by external forces continue in a state of uniform motion in a straight line.

(ii) All solid bodies expand when heated.

(iii) All swans are white.

These examples are of the form (iv).

(iv) $\forall x\ (Ax \rightarrow Bx)$ All A's are B's

Although these sentences are meaningful, they are not entailed by any finite, consistent set of observation sentences, nor, indeed, by any consistent set of sentences *An, Bn* . . . no matter what size. Since sentences of the forms (iv) and (v) are equivalent, the same is true of negations of existential generalizations.

(v) ~∃x (Ax & ~Bx) It is not the case that something is A but not B.

Fact 2: Uses of universal generalizations (and of negations of existential generalizations) are conclusively falsifiable.

The negation of an example of the form (iv) has the form (vi).

(vi) ~∀x (Ax → Bx) Not all A's are B's

Sentences of this form are equivalent to those of the form (vii).

(vii) ∃x (Ax & ~Bx) At least one A is not a B

If *A* and *B* express observable properties, then (vi) and (vii) are entailed by the set of observation sentences (viii).

(viii) An, ~Bn

Thus, uses of the corresponding universal generalizations of the form (iv), and of negations of existential generalizations (of the form (v)), are conclusively falsifiable.

Fact 3: Uses of existential generalizations (and of the negations of universal generalizations) are not conclusively falsifiable.

A use of a sentence S is conclusively falsifiable iff a corresponding use of the negation of S is conclusively verifiable. Since a use of the negation, (v), of the existential generalization, (vii), is *not* conclusively verifiable, a corresponding use of the existential generalization (vii) is *not* conclusively falsifiable. Similarly, since uses of the universal generalization (iv) are *not* conclusively verifiable, uses of its negation, (vi), are *not* conclusively falsifiable.

It follows from these facts that Attempts 1 and 2 exclude large classes of meaningful sentences. Attempt 1 wrongly characterizes many meaningful universal generalizations, and many meaningful negations of existential generalizations, as meaningless. Attempt 2 wrongly characterizes many meaningful existential generalizations, and many meaningful negations of universal generalizations, as meaningless. Both attempts also characterize certain sentences as meaningful, while denying their negations are. This result conflicts with two principles that were widely held by logical empiricists.

P1. A sentence is (cognitively) meaningful iff uses of it (in conformity with the conventions that govern it) are true or false.

P2. Uses of the negation of a sentence S are true (false) iff uses of S are false (true).

For all these reasons, Attempts 1 and 2 had to be rejected.

This brings us to the third attempt to formulate the verifiability criterion of meaning.

ATTEMPT 3

A nonanalytic, noncontradictory sentence S is empirically meaningful iff uses of S in conformity with the linguistic conventions that govern it are either conclusively verifiable or conclusively falsifiable.

When 'A' and 'B' stand for observable characteristics, this formulation handles the universal generalization ⌜All A's are B's⌝ because uses of it are conclusively falsifiable, and it handles the existential generalization ⌜At least one A is a B⌝ because uses of it are conclusively verifiable. So, both types of generalization are characterized as meaningful by Attempt 3. But three other problems remain.

The first concerns mixed quantification—sentences that contain both a universal and an existential quantifier. Here are two examples.

1. For every substance, there is a solvent. ∀x (Sx → ∃y Dxy)
2. For every man, there is a woman who loves him. ∀x (Mx → ∃y (Wy & Lyx))

Since these are universal generalizations, their uses are not conclusively verifiable. So if the sentences are meaningful, then, according to Attempt 3, their uses must be conclusively falsifiable. In order for a use of (1) to be false, a (potential) use of at least one of its instances—given in (1-Ia)—must be false; or, what is saying the same thing, a (potential) use of least one of the sentences in (1-Ib) must be true.[6] (Here we assume that we can generate a single name for each object and that the lists may be infinite.)

1-Ia. Sa → ∃y Day, Sb → ∃y Dby, Sc → ∃y Dcy, . . .
1-Ib. Sa & ∀y ~Day, Sb & ∀y ~Dby, Sc & ∀y ~Dcy, . . .

But since each conjunction in (1-Ib) has a conjunct that is a universal generalization, *none* of the conjunctions is entailed by any finite, consistent set of observation sentences. Since each conjunction is logically independent of the others, no finite, consistent set of observation sentences entails the disjunction of any pair of conjunctions, the disjunction of any trio, etc. Since a use of at least one of the disjunctions must be true if any use of (1) is to be false, no finite consistent set of observation sentences entails the negation of (1). Thus a use of (1) (in conformity with its governing conventions) isn't conclusively falsifiable. Since (1) isn't conclusively verifiable, Attempt 3 classifies it as meaningless, despite the fact that it is clearly meaningful.[7] The same reasoning applies to sentence (2).

The second problem with Attempt 3 involves other quantifications of the sort illustrated in (3) and (4).

3. There are more A's in the universe than B's.
4. Most A's are B's.

No finite, consistent set of observation sentences of the sort in (5) entails (3) or (4).

[6] For a use of *Sa* → ∃*y Day* to be false is for the corresponding use of *Sa* & ~∃*y Day* to be true. ~∃*y Day* is equivalent to ∀*y* ~*Day*.

[7] We here rely on a Moorean confidence in the meaningfulness of these sentences (plus their ubiquity in science) that exceeds our confidence in any philosophical thesis about meaning that may conflict with it.

5. Aa, Ab, Ac, . . . Bn, Bo, Bp, . . .

In order for such an entailment to exist, one would have to add to (5) some claim to the effect that the A's and B's enumerated in (5) are all there are.[8] But that sentence wouldn't be regarded by the logical empiricists as observational. Thus, uses of sentences like (3) and (4) wouldn't count as conclusively verifiable—or, by similar reasoning, conclusively falsifiable. Since such sentences are meaningful, Attempt 3 wrongly characterizes meaningful sentences of this type as meaningless.

The third difficulty with Attempt 3 plagued all attempts by logical empiricists to formulate a criterion of meaning built on the idea that a nonanalytic, noncontradictory sentence is meaningful only if the truth or falsity of uses of it could, in principle, be established by deductive reasoning from consistent sets of sentences used to record observations. This excludes much of natural science, including examples like (6).

6. The surface is being bombarded with electrons.

Scientists developing the atomic theory didn't directly observe electrons. Nor did they start with a finite, consistent set of sentences used to make observational claims, and go to their logic books to deduce (6) from that set. They also couldn't appeal to simple, enumerative induction. In short, they didn't start with observations and then deduce, or induce, (6) from them. Rather, they posited the existence of electrons as a way of explaining, and making predictions about, observable events.

The process works roughly as follows: Sentences like (6) are used, together with other sentences of one's scientific theory (often including some used to record true observations), to entail further observational sentences. If uses of all these observational consequences turn out to be true, the theory is, to that extent, confirmed. If some turn out to be false, the theory must be modified. The logical empiricists introduced the term *weak verifiability* to describe the relationship that uses of theoretical sentences like (6) stand to observational events that may confirm or disconfirm them.

How are uses of theoretical sentences assessed for truth or falsity? By itself, (6) doesn't entail any observation sentences. To get such consequences, one must combine (6) with other sentences of one's theory. Logical empiricists like Ayer wanted to say that (6) is empirically meaningful because uses of it, together with other statements, allow us to make empirical predictions we would not be in a position to make without it. They needed a new formulation of the verifiability criterion of meaning to capture this idea.

[8] One would also have to include claims asserting the nonidentity of the objects mentioned.

4. MEANINGFULNESS AS WEAK VERIFIABILITY

According to the new strategy, what makes empirical sentences meaningful is not that uses of them are, or make, statements that can be *proved* true, or false, by observations we could make. What makes them meaningful is that such observations are *relevant* to determining the truth or falsity of those statements. If including a sentence S in a theory allowed one to deduce observation sentences expressing predictions that couldn't otherwise be made, then the truth of the predictions would *support* (without conclusively establishing) the statement S is used to make, while the falsity of the predictions would disconfirm the statement (without conclusively refuting it). Since logical empiricists viewed scientific hypotheses that are confirmed or disconfirmed in this way as paradigmatic examples of uses of meaningful empirical sentences, they needed a criterion of meaning that would count those sentences as meaningful.

Here is Ayer's discussion of the matter in chapter 1 of *Language, Truth, and Logic*.

> Accordingly, we fall back on the weaker sense of verification. We say that the question that must be asked about any putative statement of fact is not, *Would any observations make its truth or falsehood logically certain?* but simply, *Would any observations be relevant to the determination of its truth or falsehood?* And it is only if a negative answer is given to this second question that we conclude that the statement under consideration is nonsensical.
>
> To make our position clearer, we may formulate it in another way. Let us call a proposition which records an actual or possible observation an experiential proposition. Then we may say that it is the mark of a genuine factual proposition, not that it should be equivalent to an experiential proposition, or any finite number of experiential propositions, but simply that some experiential propositions can be deduced from it in conjunction with certain other premises without being deducible from those other premises alone.[9]

This gives us Attempt 4.

ATTEMPT 4

A nonanalytic, noncontradictory sentence S is meaningful iff S, by itself, or in conjunction with certain further premises P, Q, R, . . . , entails some observation sentence O not entailed by P, Q, R, . . . alone.

Ayer's idea was that a sentence uses of which can play a role in explaining or predicting observations must be meaningful. The reason for the final qualifying clause is that if O were entailed by P, Q, R, . . . alone, then S would play no role in making the predictions, and uses of S would not, thereby, be connected to experience. He apparently thought that uses of

[9] Ayer (1936 [1946]), pp. 38–39.

metaphysical sentences could never be so connected, and thus that those sentences would be labeled meaningless. But, as he indicates in the introduction to the second edition of *Language, Truth, and Logic,* published in 1946, he came to realize that he was wrong.

> I say [in chapter 1 of the first edition] of this criterion that it "seems liberal enough," but in fact it is far too liberal, since it allows meaning to any statement whatsoever. For, given any statement "S" and an observation statement "O", "O" follows from "S" and "if S then O" without following from "if S then O" alone. Thus, the statements "the Absolute is lazy" and "if the Absolute is lazy, this is white" jointly entail the observation-statement "this is white," and since "this is white" does not follow from either of these premises, taken by itself, both of them satisfy my criterion of meaning. Furthermore, this would hold good for any other piece of nonsense that one cared to put, as an example, in place of "the Absolute is lazy," provided only that it had the grammatical form of an indicative sentence. But a criterion of meaning that allows such latitude as this is evidently unacceptable.[10]

As Ayer saw it, the problem arises from not putting restrictions on the supplementary premises P, Q, R, used in testing the meaningfulness of an arbitrary sentence. Since any sentence S can always be combined with the supplementary premise ⌜(S → O)⌝ to entail O, Ayer concludes that Attempt 4 accepts every sentence as meaningful. That's right, provided that ⌜(S → O)⌝ doesn't entail O by itself. Can one always assume this? Is it true that for any sentence S, one can always find a supplementary premise ⌜(S → O)⌝ that doesn't entail O by itself, and hence can be used in Attempt 4 to generate the conclusion that S is meaningful?

For all intents and purposes it is. For any *nonanalytic sentence* S, there is an observation sentence O and premise ⌜(S → O)⌝ such that O is entailed by S and ⌜(S → O)⌝ without being entailed by ⌜(S → O)⌝ alone. Consider observation sentences O1 and O2.

O1: The light is on.
O2: The light is not on (i.e., off).

The conjunction of O1 and O2 is inconsistent. Suppose that ⌜(S → O1)⌝ and ⌜(S → O2)⌝ entailed O1, and O2, respectively. If so, then ⌜(~S ∨ O1)⌝ and ⌜(~S ∨ O2)⌝ would entail O1, and O2, respectively, and ⌜~S⌝ would entail both O1 and O2.[11] Since O1 and O2 are inconsistent, this could happen only if ⌜~S⌝ were a contradiction, and S was analytic. So, for any *nonanalytic* sentence S, either S is counted meaningful by Attempt 4 because O1 is entailed by S plus ⌜(S → O1)⌝, without being entailed by ⌜(S → O1)⌝ alone,

[10] Ibid., pp. 11–12. Ayer credits the point in the passage to Isaiah Berlin (1938/39).
[11] Since ⌜(A → B)⌝ is equivalent to ⌜(~A ∨ B)⌝ and a disjunct entails any disjunction of which it is a part.

or S is counted meaningful because O2 is entailed by S plus ⌜(S → O2)⌝, without being entailed by ⌜(S → O2)⌝ alone. Since analytic sentences are meaningful, Attempt 4 does lead to the absurd result that all sentences are meaningful.

Although Ayer admitted this in the introduction to the second edition of his book, he still accepted the idea behind Attempt 4. The problem, he thought, was that it placed no restrictions on the supplementary principles used in testing S. The *reductio* seemed to arise because the sentence, ⌜(S → O)⌝, chosen to combine with S couldn't *itself* be shown to be meaningful prior to showing that S was. This suggested modifying Attempt 4 by restricting supplementary premises to those that had already been proved meaningful, *prior* to their use in testing the meaningfulness of other sentences. In presenting the new attempt, I will speak of *sentences* as observational, directly verifiable, or indirectly verifiable. An observation sentence is one a use of which, in conformity with the linguistic conventions that govern it, would predicate some observational property or relation of some observable object or objects. Directly verifiable sentences stand in certain entailment relations to observation sentences, while indirectly verifiable sentences stand in entailment relations to observation, directly verifiable, and certain other indirectly verifiable sentences.

Here is Ayer's final attempt.

ATTEMPT 5

S is *directly verifiable* iff (a) S is an observation sentence; or (b) S by itself, or in conjunction with one or more *observation sentences* P, Q, R, . . . , entails an observation sentence not entailed by P, Q, R, . . . alone.

S is *indirectly verifiable* iff (a) S, by itself, or in conjunction with other sentences P, Q, R, . . . , entails a *directly verifiable* sentence D that is not entailed by P, Q, R, . . . alone; and (b) the other sentences P, Q, R, . . . , are all either *analytic, directly verifiable*, or can be shown independently to be *indirectly verifiable*.

A nonanalytic, noncontradictory sentence S is empirically *meaningful* iff S is either directly or indirectly verifiable. (Analytic and contradictory sentences are, by definition, meaningful too.) [12]

The definition of *indirect verifiability* works in stages. At the first stage, we select a sentence and test whether it plus some directly verifiable (or analytic) sentences P, Q, R entail a directly verifiable sentence not entailed by P, Q, R, alone. Call any sentence passing this test *stage-1-indirectly-verifiable*. At stage 2 we select a new sentence S that is neither directly verifiable nor stage-1-indirectly-verifiable. We test whether S plus some P, Q, R that are either directly verifiable, stage-1-indirectly-verifiable, or analytic will entail some directly verifiable sentence not entailed by P, Q, R, alone. If S passes this test, it is *stage-2-indirectly-verifiable*. We repeat the process at stage 3,

[12] Adapted from Ayer (1936 [1946]), p. 13.

using sentences previously shown to be directly or indirectly verifiable as supplementary premises, to arrive at *stage-3-indirectly-verifiable-sentences*. The process may be repeated indefinitely many times. Any sentence passing the test at any stage counts as indirectly verifiable, and hence meaningful. But the only way a sentence can be so counted is by drawing out consequences of it in combination with sentences the meaningfulness of which has already been shown to be in accord with the criterion. Because of this, Ayer thought that he had avoided the problem that led to the collapse of Attempt 4.

Here are some examples. Let 'O_1a' and 'O_2a' be observation sentences, neither of which entails the other. (Let 'a' be a name and 'O_1x' and 'O_2x' be formulas containing a variable 'x'.) By the definition of direct verifiability, (7) and (8) are directly verifiable.

7. $(O_1a \rightarrow O_2a)$ e.g., If I drop this book, it will fall.
8. $\forall x\,(O_1x \rightarrow O_2x)$ e.g., If I drop any book, it will fall.

If 'O_3' is an observation sentence the conjunction of which with 'O_1a' doesn't entail 'O_2a', then (9) will also be directly verifiable.

9. $(O_3 \rightarrow \forall x\,(O_1x \rightarrow O_2\,x))$ e.g., If I flip the switch, then every light will go on.

When O is *any* observation sentence and DV is *any* directly verifiable sentence, $\ulcorner(O \rightarrow DV)\urcorner$ always counts as meaningful.

PROOF:
$\ulcorner(O \rightarrow DV)\urcorner$ plus O entails DV. If O alone doesn't entail DV, then $\ulcorner(O \rightarrow DV)\urcorner$ is indirectly verifiable. If O does entail DV, then $\ulcorner(O \rightarrow DV)\urcorner$ is a tautology hence analytic. Either way it counts as meaningful.

We next show that the negation of a directly verifiable sentence always counts as meaningful.

PROOF:
Let DV be any directly verifiable sentence, and let O be any observation sentence the negation of which is an observation sentence not entailed by DV—i.e., both O and $\ulcorner\sim O\urcorner$ are observational and DV doesn't entail $\ulcorner\sim O\urcorner$. For any DV, there will always be such an O.[13] We have just seen that $\ulcorner(O \rightarrow DV)\urcorner$ is always either indirectly verifiable or analytic. $\ulcorner\sim DV\urcorner$ plus $\ulcorner(O \rightarrow DV)\urcorner$ entails $\ulcorner\sim O\urcorner$. Since (by hypothesis) $\ulcorner\sim O\urcorner$ isn't entailed by DV alone, $\ulcorner\sim O\urcorner$ isn't entailed by $\ulcorner(\sim O \vee DV)\urcorner$. (Anything entailed by a disjunction is entailed by both disjuncts.) Since $\ulcorner(\sim O \vee DV)\urcorner$ is equivalent to $\ulcorner(O \rightarrow DV)\urcorner$, this means that $\ulcorner\sim O\urcorner$ isn't entailed by $\ulcorner(O \rightarrow DV)\urcorner$ alone. So, $\ulcorner\sim DV\urcorner$ is indirectly verifiable, and hence meaningful.

[13] Directly verifiable sentences are noncontradictory. So if S and $\ulcorner\sim S\urcorner$ are observational, at least one won't be entailed by DV. Whichever it turns out to be may play the role of $\ulcorner\sim O\urcorner$ in the argument.

This is good. We want ⌜~S⌝ to be meaningful when S is. We have just shown that when S is directly verifiable, Ayer's final criterion validates this.

Nevertheless, the criterion is demonstrably inadequate, as is shown by three problems—one due to Carl Hempel, one to Alonzo Church, and one inspired by Church.[14] Here is Hempel's problem. We let S be any nonanalytic, meaningful sentence that expresses a truth, and we let N be some nonsensical sentence. Ayer's final criterion counts ⌜(S&N)⌝ as meaningful, since if S is directly or indirectly verifiable, ⌜(S&N)⌝ will be the same. Ayer also holds that uses of (cognitively) meaningful sentences are either true or false. So, he must hold that a use of (S&N) is true or false. Either choice is problematic. If it is true, then a use of N must also be true, since N is entailed by (S&N). But if N is meaningless, no use of it (in accord with linguistic convention) can be true. Suppose, then, that the use of ⌜(S&N)⌝ is false. Then a use of ⌜~(S&N)⌝ must be true, in which case our use ⌜~N⌝ must be true because ⌜~N⌝ is entailed by S and ⌜~(S&N)⌝, the uses of which are both true. So, ⌜~N⌝ must be meaningful. But that is impossible since, by hypothesis, N is meaningless.

This problem is a *reductio ad absurdum* of the conjunction of Attempt 5 with the subsidiary principles P1 and P2.

P1: A sentence is (cognitively) meaningful iff its uses are either true or false.
P2: A use of ⌜~S⌝ is true (false) iff the corresponding use of S is false (true).

Whether or not Hempel's problem is a conclusive objection to Attempt 5 depends on whether or not proponents of that proposal could reasonably reject P1 or P2. Conceivably, Ayer might have been willing to give up P1 by treating ⌜(S&N)⌝ as meaningful because it entails something meaningful, while denying uses of it a truth value on the grounds that uses of N lacks truth value. But whether or not such a move is feasible is moot, since Church's problem is enough to refute Attempt 5 by itself.

In his review of the second edition of *Language, Truth, and Logic*, Church showed that for every sentence S, Ayer's final formulation of the verifiability criterion of meaning counts either S or ⌜~S⌝ as meaningful. His argument can easily be strengthened to show that the criterion classifies every sentence as meaningful.[15] Here is the argument:

S1. Let P, Q, R be observation sentences none of which entail the others.
S2. Let S be any sentence.
S3. Let (a) be the sentence ⌜(~P &Q) ∨ (R & ~S)⌝.

[14] Hempel (1950), Church (1949).
[15] This strengthening of Church's result makes implicit use, at step 10, of an assumption not used by him—namely that there are some observation sentences, the negations of which are also observation sentences. If by *an observation sentence* we mean one the truth or falsity of uses of which can be determined by simple observation, this assumption seems innocuous—think of *this is red* and *~this is red*.

S4. R is entailed by (a) plus P. Since (by hypothesis) R isn't entailed by P alone, (a) is directly verifiable.
S5. Q is entailed by (a) plus S.
S6. If Q is not entailed by (a) alone, then S is indirectly verifiable, and so is meaningful.
S7. If Q is entailed by (a) alone, then Q is entailed by its right disjunct (b): ⌜(R & ~S)⌝.
S8. If (b) does entail Q, then ⌜~S⌝ and R together entail an observation sentence Q that is not entailed by R alone—in which case ⌜~S⌝ is directly verifiable.
S9. So (from S7 and S8), if Q is entailed by (a) alone, then ⌜~S⌝ is directly verifiable.
S10. We have already shown in our discussion of Attempt 5 that the negation of a directly verifiable sentence is always indirectly verifiable, and hence meaningful. So, if ⌜~S⌝ is directly verifiable, then both ⌜~S⌝ and S are meaningful.
S11. So (from S9 and S10), if Q is entailed by (a) alone, then S is meaningful.
S10. So (from S6 and S11), if Q is, or is not, entailed by (a) alone, then S is meaningful.
S11. Since Q is always entailed by (a) alone, or not entailed by (a) alone, S is meaningful (by Ayer's criterion) no matter what S we choose.

The final problem with Attempt 5 is a variant of Church's argument put in a more revealing form. Recall the problem with Attempt 4 that motivated Attempt 5. For any nonanalytic S, there is an observation sentence O such that S plus ⌜(S → O)⌝ entails O, even though ⌜(S → O)⌝ doesn't entail O by itself. This was enough for Attempt 4 to count S as meaningful. That problem can be recreated in a nearly identical form for Attempt 5. For any nonanalytic S, there is a pair of observation sentences O and R such that S plus ⌜((S ∨ R) → O)⌝ entails O, and either (i) S counts as meaningful because ⌜((S ∨ R) → O)⌝ doesn't entail O, or (ii) S counts as meaningful because the entailment of O by ⌜((S ∨ R) → O)⌝ shows ⌜~S⌝ to be directly verifiable. In short, all the extra complexity of Attempt 5 is rendered useless when one appeals to the premise ⌜((S ∨ R) → O)⌝ rather than ⌜(S → O)⌝.

S1. Let S be any sentence.
S2. Let R and ⌜~R⌝ be incompatible observation sentences neither of which entails the observation sentence O.
S3. S plus ⌜((S ∨ R) → O)⌝ entails O.
S4. ⌜((S ∨ R) → O)⌝ is directly verifiable, because it plus R entails the observation sentence O, which is not entailed by R itself.
S5. So (from S3 and S4), if O isn't entailed by ⌜((S ∨ R) → O)⌝ alone, then S is meaningful.
S6. If O is entailed by ⌜((S ∨ R) → O)⌝ alone, then O is entailed by ⌜~(S ∨ R) ∨ O⌝ (which is equivalent to ⌜((S ∨ R) → O)⌝), in which case O is entailed

by ⌜~(S ∨ R)⌝, and hence by ⌜(~S & ~R)⌝. But that means that ⌜~S⌝ is directly verifiable, since it, plus the observation sentence ⌜~R⌝, entails the observation sentence O, which is not entailed by ⌜~R⌝ alone. So, if O is entailed by ⌜((S ∨ R) → O)⌝ alone, then ⌜~S⌝ is directly verifiable.

S7. We have already shown in our discussion of Attempt 5 that the negation of a directly verifiable statement is always indirectly verifiable, and hence meaningful. Thus, if ⌜~S⌝ is directly verifiable, then both ⌜~S⌝ and S are meaningful.

S8. So (from S6 and S7), if O is entailed by ⌜((S ∨ R) → O)⌝ alone, then S is meaningful.

S9. So (from S5 and S8), if O is, or isn't, entailed by ⌜((S ∨ R) → O)⌝ alone, then S is meaningful.

S10. Thus every sentence is meaningful (according to Ayer's criterion).

The collapse of Ayer's final formulation of the empiricist criterion was the beginning of the end of attempts to formulate the empiricist criterion of meaning in terms of either strong or weak verifiability. A few attempts were made to save the criterion from objections like those we have considered, but none proved successful. Either obviously meaningful sentences of science were wrongly characterized as meaningless, or obviously meaningless sentences were classified as meaningful. Hence another approach was needed.

5. EMPIRICAL MEANINGFULNESS AS TRANSLATABILITY INTO AN EMPIRICIST LANGUAGE

By the late 1940s and early '50s there were still philosophers who thought that there was something valuable in the idea of somehow linking empirical meaning to empirical observation. Carl Hempel, who was among the critics of standard formulations of the verifiability criterion of meaning, was one of them. In his article "The Empiricist Criterion of Meaning," first published in 1950, he catalogs the failures of the logical empiricists to come up with successful formulations of their criterion in terms of either strong or weak verifiability. He then considers a different approach, which might be called *the translatability criterion of meaning*.

THE TRANSLATABILITY CRITERION OF MEANING

A sentence is empirically meaningful iff it can be translated into an empiricist language—i.e., iff it can be translated into a version of Russell's language of *Principia Mathematica* in which the only predicates are those expressing observable properties, plus predicates definable from them together with the truth-functional operators and quantifiers of Russell's language.

Neither this criterion, nor the others he discusses, was original with Hempel. The translatability criterion was drawn from Carnap (1936/37).

Although Hempel doesn't endorse this criterion, he cites four of its virtues. First, it makes explicit provision for universal and existential quantifications. Since the Russellian language includes both quantifiers, sentences containing them aren't excluded on principle from being meaningful, as they were by criteria based on conclusive verifiability and conclusive falsifiability. Second, Hempel assumes, plausibly, that sentences like *The absolute is perfect* can't be translated into an empiricist language. So, the new criterion does not, as Ayer's later criteria did, end up counting all sentences as meaningful. Third, since *The absolute is perfect* can't be translated into an empiricist language, no meaningful conjunctions or disjunctions can contain it as a constituent. Fourth, the translatability criterion captures the idea that if S is meaningful, its negation is too, since if the translation of S is P, then the translation of the negation of S will also be translatable.

Hempel also notes two serious problems with the translatability criterion. The first involved what he called *disposition terms*, which he characterized as "*terms which reflect the disposition of one or more objects to react in a determinate way under specified conditions.*"[16] He cites, *temperature, electrically charged, magnetic, intelligent,* and *electrical resistance* as examples of such terms. This list is somewhat surprising. A clear example of a disposition term is *fragile*, which means, roughly, *is disposed to break when struck*. But it hardly seems that Hempel's example, *temperature*, means *is disposed to* ν, for any choice of 'ν'. Still, what he had in mind is clear enough. Consider *the temperature of x is 90 degrees Fahrenheit*. Hempel doesn't regard this as a simple observation sentence—presumably because ordinary observation, unaided by special measuring devices, and unmediated by any background theory containing non-observational terms, isn't enough to determine whether uses of it are true. So, he thinks, it is translatable into an empiricist language only if the relational two-place predicate *the temperature of x = y* can be completely defined in terms that are purely observational.

Here are two possible definitions:

D1. For any object x and number y, the temperature of x = y degrees Fahrenheit iff x is in contact with a thermometer that measures y degrees Fahrenheit on its scale.

D2. For any object x and number y, the temperature of x = y degrees Fahrenheit iff (x is in contact with a thermometer → the thermometer it is in contact with measures y degrees Fahrenheit on its scale).

D1 is obviously inadequate because it wrongly characterizes any object not in contact with a thermometer as not having any temperature. D2 is similarly inadequate because it wrongly characterizes any object not in contact with a thermometer as having every temperature. (The right side of D2 is a material conditional, which is equivalent to the disjunction of

[16] (1950 [1959]), p. 119.

its consequent and the negation of its antecedent.) Hempel notes that we might have more success if we allowed the use of counterfactual conditionals, as in D3.

> D3. For any object x and number y, the temperature of x = y degrees Fahrenheit iff (if it were the case that x was in contact with a thermometer then the thermometer would measure y degrees Fahrenheit on its scale).

However, since counterfactual conditionals are *not* truth-functional, and so not part of Russell's language, D3 is not available in Hempel's "empiricist language."

Might we liberalize the criterion by allowing empiricist languages to include counterfactual conditionals like D3? Hempel says, "*This suggestion would provide an answer to the problem of defining disposition terms if it were not for the fact that no entirely satisfactory account of the exact meaning of counterfactual conditionals seems to be available at present.*"[17] Although this comment was true when written, it is arguably not so today. By the late 1960s and early '70s, several philosophers, including most prominently Robert Stalnaker and David Lewis, had adapted the framework of possible worlds semantics developed by Rudolf Carnap, Saul Kripke, Richard Montague, and others to the study of counterfactual constructions.[18] Roughly put, *If A had been the case, then B would have been the case* is true at a possible state of the world w iff among the world-states at which A is true, some at which B is true are more similar to w than any at which B is false. More informally, *If A had been the case, then B would have been the case* is true at w iff a world-state differing from w in the minimum amount needed to make A true is one at which B is true. This approach is now widely accepted.

Since this development renders Hempel's critical comment outdated, one might naturally ask whether allowing definitions like D3 into empiricist languages would solve the problems posed for the translatability criterion of meaning by notions like *temperature*. There are two reasons to think not. First, the semantic apparatus drawn from possible worlds semantics to explain counterfactuals contains elements that would have been regarded with suspicion by at least some proponents of the empiricist criterion of meaning. The notion of a possible state of the world is now standardly understood as a *metaphysical* notion of possibility that isn't reducible to, or explainable in terms of, purely linguistic conceptions of possibility, necessity, or analyticity. Hence, using it to characterize an empiricist language might well be viewed by logical empiricists as importing metaphysics into a criterion of meaning designed to exclude metaphysics as meaningless. To put the matter more dramatically, the semantic developments that gave us a logic of counterfactuals can't naturally be used to

[17] Ibid., page 120.
[18] Stalnaker (1968, 1975), Lewis (1973), Carnap (1956), Kripke (1959, 1963), Montague (1974).

save logical empiricism because those developments were based on the presupposition that the logical empiricists were wrong about meaning in general, and possibility in particular.

The second reason for thinking that definitions like D3 don't solve the problems posed by terms like *temperature* for the translatability criterion of meaning is more prosaic. If definitions like this are noncircular, then they won't cover all the cases, and so will fail as definitions. To see this, it suffices to note that some things are very hot; for example, the temperature of the sun is so high that a thermometer put up against it would melt or explode, and not give any reading. Nevertheless, the sun has a temperature. Since D3 does not allow for this, it is not an adequate definition.

One might object to this conclusion by saying that D3 is incorrect only if we take the word *thermometer* to mean the sort of ordinary existing thermometers with which we are familiar. Surely, the objector might continue, we can *imagine* thermometers that wouldn't melt or explode, even on the sun. If we take the word *thermometer* in D3 to be talking about them, then the counterexample disappears. Let's see. Suppose we use the word *thermometer* in D3 to cover these nonexistent but conceivable devices. What, then, are we taking *thermometer* to mean? A natural thought, I suppose, is that *thermometer* means *a device (however constructed) for accurately measuring temperature*. If so, then it may be true that if n is the temperature of the sun, and if a thermometer—i.e., an accurate device for measuring the temperature of the sun—were placed on the sun, then the device would read n on its scale. But the cost of saving D3 from this counterexample has been to define *thermometer* in terms of the antecedently understood notion of *temperature*, rather than the other way around. When so understood, D3 isn't a definition of temperature. So we still haven't succeeded in rendering statements about temperature translatable into an empiricist language. Consequently, the problem for the translatability criterion of meaning remains.

Another defect with the criterion mentioned by Hempel involves what he calls *theoretical constructs*, examples of which include the terms *electron, gravitational potential,* and *electric field*. As Hempel defined an empiricist language, the only predicates allowed are observation predicates, and predicates that can be defined in terms of observation predicates plus Russell's logical apparatus. Hempel notes that a predicate like *is an electron* is neither observational, nor definable in strictly observational terms. Since this means that it would be excluded from an empiricist language, the translatability criterion of meaning wrongly characterizes sentences about electrons and other theoretical entities as meaningless.

Hempel took this problem to show that empiricists must shift the focus of their criterion of meaning away from individual sentences, and toward systems of sentences. According to him, what makes sentences about theoretical entities meaningful is that they are embedded in a network of observational and non-observational sentences that can be used to make testable predictions. These predictions are the product of the different

aspects of the system working together. So, if one is given a set of observational predictions made by using a theory, one cannot, generally, match up each prediction with an isolated hypothesis expressed using a single sentence of the theory. Hempel suggests that this is the crucial fact that makes it impossible to define theoretical terms in isolation. If for each statement made using a sentence S involving a theoretical term, we could isolate a set of predictions made by one's use of S alone, and if those predictions exhausted the contribution made by uses of S to the predictions derived using the theory as a whole, then we could simply identify the meaning of S with those predictions. However, the interdependence of S with other sentences in the theory makes this impossible. Thus, what we have to look for is not the empirical content of each individual use of a sentence taken in isolation, but rather the role of each such use in the use of an articulated system which, as a whole, has empirical content.

6. LESSONS

What, then, is left of the empiricist criterion of meaning? In effect, it evolved into the claim that a nonanalytic, noncontradictory sentence is meaningful when uses of it play a functional role in some larger system that is used to make observational predictions. Although that sounds reasonable, it is vague and open-ended. What counts as a theoretical system? What is the empirical meaning or content of such a system? What role must the use of an individual sentence play in the system in order to be counted as meaningful in virtue of its contribution to the meaning of the whole? Are only systems that are actually used capable of conferring meaning on a sentence, or may a sentence be meaningful because it is *conceivable* that it should play an appropriate role in some merely possible systems? None of these questions are seriously addressed by Hempel, let alone answered. Still, the shift in emphasis away from the individual sentence to the system or theory is significant. The key notion is that the system as a whole is the thing that has observational consequences. So, if meaning is still to be analyzed in terms of such consequences, the natural units of meaning—the things to which empirical criteria of meaningfulness apply—should be entire theories or systems, rather than individual sentences.

This move toward *linguistic holism* was one of two major responses that grew out of the history of failed attempts to formulate a verificationist theory of meaning for individual sentences. Quine, whose philosophy will be examined in volume 3, was the chief proponent of this response. He argued that meaning really is explainable in terms of verification on the basis of observational consequences, but since these consequences can't be portioned out over sentences taken individually, entire theories are the primary bearers of meaning, or content. The other main historical reaction to the failure of logical empiricism rejected the idea that meaning can be

understood, or analyzed, in terms of verification, and attempted to find another way of understanding it. Following the later Wittgenstein, many British philosophers in the post-positivist period—John L. Austin, Gilbert Ryle, Peter Strawson, Richard M. Hare, and others—attempted to explain meaning by appealing to the many different ways in which expressions are used in ordinary language, and to draw philosophical lessons from this approach. These ideas will be examined in volume 4.

Before leaving the story of unsuccessful attempts to formulate an acceptable version of the empiricist criterion of meaning, there is one further philosophical lesson, of a broadly Moorean sort, to be drawn. In discussing Moore's response to skepticism in volume 1 (Soames 2014), I emphasized one of his methodological points about philosophical theories of knowledge. He contended that no matter how attractive a theory of the necessary and sufficient conditions for knowledge might initially seem, it must be tested against the mass of our most confident commonsense judgments about what we know, and what we don't. So, if any philosophical theory of knowledge can be shown to conflict with most of what we ordinarily take ourselves to know, then that theory, rather than the commonsense judgments, will be suspect. The same point can be made about theories of meaning. Even though the logical empiricists had an initially attractive theory about what empirical meaning must be, the fact that different formulations of it repeatedly conflicted with our most confident pretheoretic judgments was, correctly, taken to cast doubt on the philosophical theory, rather than on the mass of our pretheoretic judgments.

The general point extends well beyond the particular theories developed by the logical empiricists. Any theory of meaning we might construct, any theory of the form

S is meaningful iff . . .

must be answerable, to some considerable extent, to our pre-philosophical judgments of what is meaningful and what isn't. This is true no matter whether the theory is aimed at *describing* our ordinary concept of meaning, or whether its aim is the partially *revisionary* one of modifying our ordinary concept by purging it of obscure or problematic elements in order to solve theoretical problems. Verificationist theories of meaning were consciously *reformist*. The logical empiricists thought it was a virtue of their theories that they were not completely faithful to every confident judgment about meaning that we ordinarily make. In some cases, they may have been right. But we have seen that even theorists aiming at substantial reform can't afford to stray too far from our ordinary, pre-philosophical judgments. As one goes farther down the reformist path, the implausible consequences of one's theory will threaten to outweigh its initial attractiveness. This is not to say that no philosophical revision of our ordinary judgments, or of our ordinary pre-philosophical concepts, can be justified. It is to say that our pre-philosophical judgments constrain even philosophically well-motivated theories.

Part Three

IS ETHICS POSSIBLE?

CHAPTER 12

Ethics as Science

1. Schlick's Vision of Ethics
2. Egoism, Altruism, and Morality
3. Paths to Virtue
4. Might There Be a Scientific Basis for Morality?

1. SCHLICK'S VISION OF ETHICS

In 1930 Schlick published a book that was translated into English in 1939 as *Problems of Ethics.* In it he argued that ethics should be regarded as an empirical science. Although the book received little attention in its time, and less thereafter, it deserves a fuller hearing—in part as a historical corrective of the widespread misconception that the central doctrines of the logical empiricists left no room for normative theory or non-cognitivist metaethics, and in part for its insight into how an empirical, and ultimately scientific, theory of human nature might impact ethical theory.

Schlick announces his conception of ethics as a science in the opening pages *Problems of Ethics*.

> If there are ethical questions which have meaning, and therefore are capable of being answered, then ethics is a science. For the correct answers to its questions will constitute a system of true propositions, and a system of true propositions concerning an object is the "science" of that object. Thus, ethics is a system of knowledge and nothing else; its only goal is the truth.[1]

This striking declaration that the aim of ethics is to advance our knowledge of moral truths by discovering moral facts reminds one of the early pages of G. E. Moore's 1903 classic, *Principia Ethica,* in which he announces his intention of laying the foundation of an autonomous *scientific system of ethics* independent of empirical theories and metaphysical doctrines. In

[1] Schlick (1930 [1939]), p. 1.

addition to speaking of this "science" in sections 3, 4, 5, and 14 of chapter 1, Moore describes his scientific aim in the preface.

> I have endeavored to write "Prolegomena to any future ethics that can possibly pretend to be scientific." In other words, I have endeavored to discover what are the fundamental principles of ethical reasoning; and the establishment of these principles rather than any conclusion which may be attained by their use, may be regarded as my main object.[2]

The foundational truths of Moore's envisaged normative science consisted of statements about intrinsic value of states of affairs, from which the right-making features of actions were to be derived. He took these truths to be synthetic, necessary, and knowable a priori.

Three decades later, in the heyday of logical empiricism, such statements were deemed incoherent, along with all moral theories of traditional philosophy. According to logical empiricism, there are two kinds of truth—analytic and empirical. Since the former are true solely in virtue of meaning, entirely independent of facts, any conception of ethical theory as an attempt to extend our knowledge of ethical facts must construe ethical truths as contingent and knowable only a posteriori, on the basis of ordinary observation. Schlick was the only important logical empiricist who thought that there were such truths. The others declared ethical sentences to be cognitively meaningless expressions of emotion, misleadingly packaged to look like descriptions of genuine facts.

Unlike Ayer, Carnap, and Stevenson, who sought to *replace* ethics with metaethics, Schlick took metaethics to be merely preliminary to true ethical theory, which he believed to be part of empirical psychology. Thus, he viewed his book as primarily a contribution to psychology. It was, he thought, philosophical only to the extent that—in addition to contributing to the substance of a genuine science—it also contained philosophical analyses of the needed scientific concepts. For example, while he seemed to regard fundamental ethical terms like 'good' and 'ought' as primitive and indefinable, he also thought that facts about their extensions can be discerned from regularities in their use, just as facts about the extensions of indefinable color words like 'green' and 'red' are discerned from regularities in their use.

> It is very dangerous to withdraw from this task [of formulating the concept of moral good] under the pretext that the word "good" is one of those whose meaning is simple and unanalyzable. What is demanded here need not be a definition in the strictest sense of the word. It is sufficient to indicate how we can get the content of the concept; to state what must be done in order to become acquainted with its content. It is, strictly speaking, also impossible to define what the word "green" means, but we can nevertheless fix its meaning

[2] Moore (1903), preface, p. v.

> unambiguously, for example, by saying it is the color of a summer meadow, or by pointing to the foliage of a tree. . . . In the same way, in ethics we must be able to give the exact conditions under which the word "good" is applied, even though its fundamental concept be indefinable.[3]

How are we to determine the application of fundamental ethical terms? Since ethical behavior is conduct we demand from others and ourselves, Schlick takes it to be conduct *that we fundamentally desire* which relates us to others, and them to us. This, for him, is a rock-bottom empirical fact. If our most basic desires of this type can be identified, there is no further question of justification to be raised. It is nonsense to ask, *Is what we most fundamentally value really valuable?* In the end we simply value what we do. He wants to know *what* conduct we really do value and *why* we value it. These are empirical questions about the psychological makeup of human beings.

Which human beings? Schlick realizes that different standards of conduct might be demanded by different individuals, groups, and societies in different circumstances at different times. Whether their seemingly diverse moral codes reflect genuine differences in underlying value, as opposed to varying factual circumstances, differing ranges of available actions, and different opinions about the effects of different actions, cannot, he thinks, be decided in advance. Nevertheless, he is confident that a great deal of common ground can be found. He writes:

> Whether there is actually among men a multiplicity of moralities incompatible with one another, or whether the differences in the moral world are only specious, so that the philosopher would find everywhere, under many different disguises and masks of morality, one and the same face of the one Good, we cannot now decide. In any case there are wide regions in which the unanimity and security of moral judgments is substantiated. The modes of behavior which we group together under the names reliability, helpfulness, sociability are everywhere judged to be "good," while, for example, thievery, murder, quarrelsomeness pass for "evil" so unanimously that here the question of the common property can be answered with practically universal validity.[4]

Schlick imagines the study of the morality of a given group as issuing in a hierarchical system of the norms specifying morally good conduct demanded in various circumstances. The claim that something is such a group norm is a factual claim about the conduct its members expect and demand. The enumeration of these norms is, according to him, "nothing but the determination of the concept of the good, which ethics undertakes to understand."[5]

[3] Ibid., pp. 8–9.
[4] Ibid., pp. 13–14.
[5] Ibid., p. 15.

> The determination would proceed by seeking ever new groups of acts that are recognized to be good, and showing for each of them the rule or norm which all of their members satisfy. The different norms, so obtained, would then be compared, and one would order them into new classes such that the individual norms of each class had something in common, and thus would all be subsumed under a higher, that is, more general, norm. With this higher norm the same procedure would be repeated, and so on, until, in a perfect case, one would at last reach a highest, most general rule that included all others as special cases, and would be applicable to every instance of human conduct.[6]

Schlick is well aware that there is no guarantee that a single all-encompassing moral principle will emerge, and hence that there may turn out to be a single morality with mutually independent principles, or, perhaps, multiple somewhat differing moralities governing different subgroups of agents. When moral systems are understood in this way, one justifies the claim that certain acts are morally good by citing the norm under which they fall, while justifying lower-level norms in terms of higher level norms. The process ends with the highest norm, or norms, for which no further justification makes sense.

> The question regarding the validity of a valuation amounts to asking for a higher acknowledged norm under which the value falls, and this is a question of *fact*. The question of the justification of the highest norms or the ultimate values is senseless, because there is nothing higher to which these could be referred. . . . Such norms as are recognized as the ultimate norms, or highest values, must be derived from human nature and life as facts.[7]

So far, it may seem that the science of ethics envisioned by Schlick is little more than relativistic social psychology. One extracts the codes of conduct adhered to in different societies or social groups, and one explains how internal moral justification proceeds in each, while insisting on the absurdity of imagining that there can be an external standard of justification. But this is only the first step in Schlick's envisioned moral science. At the next step, we attempt to find higher, non-moral norms that explain the ethical norms.

> It might be that the *moral* good could be shown to be a special case of a more general kind of good. . . . If [so] . . . then the question, "Why is moral behavior good?" can be answered by "Because it is good in a more general sense of the word. The highest moral norm would be justified by means of an extra-moral norm; the moral principle would be referred back to a higher principle of life.[8]

[6] Ibid., p. 15.
[7] Ibid., p. 18.
[8] Ibid., p. 24.

This may seem perplexing. Why, one may wonder, is reduction to a higher non-moral norm, or set of norms, supposed to help? If we end up with a highest norm that can't be further justified, why is it an advance if the norm is non-moral? Though Schlick's answer is less clear than one might hope, it points to the interesting path that he follows in the book.

> It is not the norms, principles, or values themselves that stand in need of and are capable of explanation but rather the actual facts from which they are abstracted. These facts are the acts of giving rules, of valuation, of approbation in human consciousness; they are thus the real events in the life of the soul. . . . And here lies the proper task of ethics. Here are the remarkable facts which excite philosophic wonder, and whose explanation has always been the final goal of ethical inquiry. That man actually approves certain actions, declares certain dispositions to be "good," appears not at all self-explanatory to the philosopher, but often very astonishing, and he therefore asks his "Why?" Now in all of the natural sciences every explanation can be conceived as a *causal* explanation. . . . therefore the "why" has the sense of a question concerning the *cause* of that psychological process in which man makes a valuation, establishes a moral claim.[9]

Earlier I mentioned that, for Schlick, morality is a system of *demands* we place not only on others, but also on ourselves. These demands are often inconvenient, bothersome, onerous, or worse. Why do we make them? It is easy to understand why we want to constrain others. But why are we willing to constrain ourselves? In part because we need the cooperation of others and can only secure it by being perceived as conforming to the rules to which we expect others to conform. Schlick recognizes this, but he doesn't take this instrumental explanation to be the whole answer.[10] If it were, we would feel perfectly fine about cheating or free-riding when not detected. In fact, however, we typically don't. Thus, something further must be going on. What is it? The answer can only be that we somehow find value in living up to our most fundamental norms. Schlick's overriding goal is to find out what we value, why we value it, and how the value, or values, in question provide the foundation for both the happiness and the moral virtue of the individual. He is convinced that the answer must come from a deep empirical study of human psychology and human nature.

So understood, Schlick's envisioned empirical science of ethics has two parts. The first part describes our actual moral norms. The second explains what it is about we human beings that makes us approve of those norms, that gives us reason to conform our actions to them, and that increases our prospects of happiness while also leading us to virtue.

[9] Ibid., pp. 24–25.
[10] See the discussion of Hobbes in Schlick (1930 [1939]), pp. 162–65.

> [T]he *determination* of the contents of the concepts of good and evil is made by the use of moral principles and a system of norms and affords a relative justification of the lower-order moral rules by the higher; scientific *knowledge* of the good, on the other hand, does not concern norms, but refers to the cause, concerns not the justification, but the explanation of moral judgments. The theory of norms asks "*What* does actually serve as the standard of conduct?" Explanatory ethics, however, asks "*Why* does it serve as the standard of conduct?[11]

If we can answer these questions we will know what our morality demands, why we wish to be moral, and the degree to which the wellsprings of morality are fixed and unchanging as opposed to reactions to transient and changeable circumstances.

2. EGOISM, ALTRUISM, AND MORALITY

Schlick's characterization of morality begins with his explication of the universal condemnation of egoism.

> Egoistic volition is for us the example of immoral volition, volition that is condemned. To condemn an act means always to desire that it should not occur. And the desire that something should not happen means . . . that the idea of its happening is unpleasant.[12]

According to Schlick, desiring something is a cognitive state in which contemplating it is combined with a positive or negative emotional charge. He takes it to be an empirical law that "deciding what to do" is a matter of allowing the intensities of these positive and negative contemplations to interact with one another until a non-neutral balance for or against some course of action is reached. The resulting decision to act is the initial cognitive stage of the action. Needing terms to label the positive and negative emotive components of desires, he adopts the awkward strategy of calling the positive cognitions "pleasant" and the negative cognitions "painful"—while admonishing us not to take these terms as standing only for familiar bodily pleasures or pains.

He applies this framework in his discussion of our disapprobation of egoism.

> Egoistic volition is for us the example of immoral volition, volition that is condemned. To condemn an act always means to desire that it should not occur. And the desire that something should not happen means . . . that the idea of its happening is unpleasant. Thus, when we ask, "Why do I condemn egoistic

[11] Ibid., p. 25.
[12] Ibid., pp. 76–77.

> behavior?", the question is identical in meaning with "Why does the idea of such behavior cause me pain?" . . . It is "Because the selfishness of another actually causes me pain directly." For its essence [the essence of egoism] is just inconsiderateness with respect to interests of fellow men, the pursuit of personal ends at the cost of those of others. But since I belong among these others, I am in danger of suffering a restriction of my joys and an increase of my sorrows at the hands of the egoist, at least in so far as his conduct impinges on my sphere of life. Where this is not the case it affects at least the feelings and lives of our fellow men, and I share in these by virtue of my social impulses; because of them I feel as my own pain the damage done to others by the egoist. Each member of human society will, on an average, react to egoism with the same feelings for the same reasons. The blame and condemnation with which they oppose it *is* nothing but *moral* censure, *moral* condemnation.[13]

It is significant that the threats Schlick recognizes as being posed to us by the egoism of others are *not* limited to our purely self-regarding goals or interests. On the contrary, he assumes that each of us has social impulses that bind us to others in a way that makes the contemplation of damage to them painful to us. This helps him explain why we condemn the egoistic conduct of others, why we often avoid behaving egoistically ourselves, and why, when we are tempted into such behavior, we may come to feel guilty about, and thus be willing to sincerely condemn, our own egoistic conduct. Because behaving egoistically threatens our own happiness by thwarting our social impulses, we have a reason, rooted in our own desires, to avoid it. For Schlick, this mix of self- and other-regarding interests is the source of our moral condemnation of egoism.

His aim is to explain all moral condemnation in the same way. Since we call conduct we morally condemn 'wrong' or 'morally bad', while calling conduct of which we morally approve 'morally good' or 'right', it would seem that the explanations he seeks should uncover the values, dispositions, and cognitive attitudes that fix the extensions of our moral terms. However, his pursuit of these extensions in chapter 4, "What Is the Meaning of 'Moral'?," takes him down a dubious path, where he argues for three conclusions. He says:

> (1) The meaning of the word "good" (that is, what is considered as moral) is determined by the opinion of society, which is the lawgiver formulating moral demands. . . .
> (2) The content of the concept "good" is determined in such a way by the society that all and only those modes of behavior are subsumed under it which society believes are advantageous to its welfare. . . .
> (3) The moral demands are established by society *only because* the fulfillment of these demands appears to be useful to it. . . . [T]he *material* meaning

[13] Ibid., p. 77.

> of the word "moral" *exhausts itself* in denoting what, according to the prevailing opinion in society, is advantageous (its *formal* meaning consists in being demanded by society).[14]

Juxtaposing this passage with the one previously cited, we see Schlick pulled in two directions. The previous passage tells us that *each one of us* has a reason—of our own—to morally demand, as it were, non-egoistic behavior of ourselves and others. This suggests that, in general, the claim that one morally ought to behave in a certain way will, if true, provide one with a reason, *grounded in one's own desires*, for so acting. However, in the second, three-part, passage the value to the individual of acting morally seems to disappear. Instead, the source of moral demands is said to lie in the *beliefs* of *society* about what is useful to *society.* Schlick is driven to this by the fact that society does make what it deems to be moral demands on its members, and also by his implicit thought that the extension of a simple indefinable term like 'good' (in its moral use) is fixed by the community's application of it to certain things and not others. The tension between the two passages arises from the possibility that *most individuals* in a given society might well *believe* that certain conduct is most beneficial to the society, even if there is no genuine reason, grounded in their own desires, for *many* in the society to act in that way. Is it really true, in such a case, that these individuals morally ought to act in the socially prescribed way?

The problem for Schlick is sharpened in his criticism of Kant.

> Kant showed correctly that the moral precepts have the character of demands, and that each appears to us as an "ought." But he could not bring himself to leave its empirical meaning to this word, in which alone it is actually used. Everyone knows this meaning: "I ought to do something" never means anything but "Someone wants me to do it." And in fact the desire of another, directed on me, is described as an ought only when that person is able to add pressure to his desire and thus to reward fulfillment and to punish neglect, or at least point out the natural consequences of observance or neglect. . . . We call such a desire a command (imperative); therefore it is of the essence of the imperative to be hypothetical, that is, to presuppose some sanction, a promise or a threat.
>
> According to our own view . . . the lawgiver who sanctions the moral commands is human society, which is furnished with the necessary power to command. Thus we may rightly say that morality makes demands on men, that they *ought* to behave in certain ways. . . . But . . . Kant cannot be satisfied with this. No matter whom he might find to be the source of the ethical command it would always be hypothetical, dependent upon the power and desire of this being, ceasing upon his absence or with a change in his desire. . . . But we have seen that a relationship to a power that expresses its desires is *essential* to the

[14] Ibid., pp. 96–97.

> concept of the ought, just as essential as the relationship to some conditions (sanctions) is for the concept of the imperative.[15]

Schlick's opposition to Kant is not in question here. What is in question in his sketch of the connection between the demands and precepts of morality, on the one hand, and the demands of a given society, together with the social rewards and punishments that go with them, on the other. In passages like this—which represent one strain of his thought—Schlick seems to be thoroughly relativist, and unable to provide any basis in which members of a society might critique its moral demands on them. However, this was not his last word.

3. PATHS TO VIRTUE

In the final chapter of *Problems of Ethics*, Schlick focuses on society's attempt—through instruction, suggestion, admonition, reward, and punishment—to socialize the individual into internalizing its moral demands. Midway through the chapter, he notes the central problem: the possibility that the socially mandated way of life may not be experienced as valuable by those who adopt it. He labels this a discrepancy between *motive pleasure* and *realization pleasure,* which leads him to a limiting principle.

> If we wish to generate lasting dependable dispositions in a person, we must take care that the realization pleasure contains what the motive pleasure promises. . . . We can now say that there is only *one* way to create motives of conduct which will prevail against all influences: and this is by *reference to actual happy consequences*.[16]

Schlick concludes that when a socially imposed moral code systematically contradicts fundamental values of large numbers of individuals, and thereby thwarts their happiness, it will not survive, but will instead be modified or transformed.

This leads him to a fundamental question.

> Thus we are confronted by the question: are the ends commended to us in the moral precepts [imposed by society] really genuine values for the individual, or do they consist in the feelings of pleasure with which society has been clever enough to equip the ideas of the ends desired by itself? We are confronted by the ancient problem: does virtue lead to happiness?[17]

Note Schlick's framing of the question—not *Does what socially passes for virtue lead to happiness*? but *Does virtue lead to happiness?* These will be the

[15] Ibid., pp. 110–12.
[16] Ibid., pp. 179–80.
[17] Ibid., p. 182.

same, if the moral demands on one are, by definition, the demands sanctioned by the conception of morality endorsed by one's society. Up to now Schlick has written as if the moral demands were the socially sanctioned demands. Now a new possibility emerges. Suppose the path to happiness is one that offers the best prospect of fulfilling our most fundamental natural impulses, including those that relate us to other people. Suppose further that the moral life mandated by one's society is *not* the path to happiness. Should we conclude that virtue and happiness diverge? Or should we instead conclude that socially mandated virtue is not true virtue, because society's conception of morality can, and often does, deviate from true morality. Schlick explores the latter prospect at the end of his book.

He begins by characterizing the social impulses as those the fulfillment of which are of paramount importance to our happiness.

> I have no doubt that experience indicates very clearly that the *social* impulses are those which best assure their bearers of a joyful life. The social impulses are those dispositions of a person by virtue of which the idea of a pleasant or unpleasant state of *another* person is itself a pleasant or unpleasant experience. . . . The natural effect of these inclinations is that their bearer establishes the joyful states of others as ends of his conduct. . . . In itself (that is apart from consequences) what an impulse is directed toward is quite indifferent to the resultant joy, and there is not the least essential reason why, for example, the pleasure in filling one's stomach should be in any way distinguished from the joy one has looking into eyes shining with happiness. The latter joy may be more difficult to understand in biological-genetic terms, but this, above all, concerns neither the philosopher nor the psychologist.[18]

Schlick's suggestion is arresting. The social impulses that lead us to be concerned with the welfare of others are as basic to our nature as the impulse to eat when hungry, while being more important for achieving happiness than any of our other natural impulses. Suppose he is right. The natural next step is to explore the possibility that our social impulses form the biological and psychological basis for genuine morality, over and above what is socially dictated. Schlick was clearly moving in that direction. However, the last sentence may be a bit of a false note. If the social impulses are as ethically important as he thinks they are, then investigating their possible biological or genetic origin should be of interest to the moral philosopher—particularly one who maintains that ethical claims are empirically verifiable or falsifiable. Perhaps Schlick is here merely registering that not knowing their ultimate causal origins is no reason to doubt their existence.

This thought is confirmed by several impressionistic passages in the next few pages about the social impulses and their connection with happiness. Here is one.

[18] Ibid., pp. 186–87.

> The social impulses constitute a truly ingenious means of multiplying the feelings of pleasure; for the man who feels the pleasure of his fellow men to be the source of his own pleasure thereby increases his joys with the increase in theirs, he shares their happiness. . . . The objection that social feelings have as a consequence the sharing of sorrow is partially justified, but does not weigh so heavily, because suffering too gives scope to the satisfaction of the social impulses, in that one can work for its alleviation.[19]

Schlick's point is only the beginning. To it we might add a number of related points. One of these is the temporal observation that while our capacity for deriving satisfaction from purely self-regarding desires often declines with age and the recognition of our own increasingly evident mortality, this is not true of our capacity for deriving satisfaction from the welfare of others and the contributions we are able to make to it. For this reason, it is natural that the attachment we feel to life as we grow older, and the enjoyment we derive from it, comes to depend more and more on our interest in, and commitments to, colleagues, friends, children, loved ones, and indeed to all who do, or will, participate in the enduring human projects with which we identify.

Schlick's next step is to connect his appreciation of the power of what he calls our "social impulses" to the possibility of a more objective conception of morality.

> Thus far in answer to the question: what paths lead to the highest values? we have discovered at least that the guide to them is to be found in the social impulses. . . . But to the concept of morality, which we investigated in chapter IV [discussed above in section 2] there is joined an indefiniteness of no small degree. We found that those dispositions are called moral which human society *believes* are most advantageous to its general welfare. Hence the content of the concept depends not only upon the actual living conditions of society, but also upon the intelligence of the class which determines public opinion, and upon the richness of its experience. This confusion and relativity is unavoidable. . . . But it remains unsatisfying that the definition of morality by the *opinion* of society makes meaningless a question which the philosopher (here becoming a moralist) would very much like to ask: namely whether what society holds to be moral really *is* so.[20]

Here, Schlick confronts the tension noted in section 2 between what initially appeared to be his resolute moral relativism and his identification of elements in human nature that might provide an objective basis for morality. His framing of the issue in this passage seems to betray a continuing uncertainty. Although he is uncomfortable with the subjectivity and

[19] Ibid., p. 189.
[20] Ibid., pp. 195–96.

relativity inherent in letting the opinion of society *determine* the content of morality, thereby rendering *the truth* of its moral judgments unchallengeable, he is also uncomfortable about assuming the role of a *moralist* by criticizing the morality of society, perhaps because he sees adopting such a role as replacing objectivity with advocacy. But what exactly is the worry? If the social impulses in human nature are as important as he thinks they are, why can't they be the source of genuine moral demands the flouting of which typically risk diminishing one's prospects for lasting happiness?

Schlick gingerly approaches this idea in the following passage by noting that, for him, the ground for morality is not a concern for "the greatest happiness for the greatest number" but the individual's own prospect of happiness. He says:

> We did not begin by seeking the causes of "moral" dispositions [i.e., those commonly so labeled], but sought the disposition which is most valuable for the agent himself, which, that is, leads with the greatest probability to his happiness. . . . [T]hus we eliminated any reference to the opinions of society. . . . And, thus, the otherwise disturbing relativity and confusion is removed in so far as this is at all possible. . . . With the most sharply defined question there would, perhaps, be given the possibility subsequently of speaking of a standard of morality, and of judging whether different moral views correspond to it or not. The philosopher could, for his purposes, *define* as moral that behavior by means of which an individual furthered his capacity for happiness, and could designate the precepts of society as "truly" moral if this criterion fitted them. . . . The formulation of a "moral principle," too, would be possible on this basis; and it would run, "At all times be fit for happiness."[21]

The idea behind this non-subjective conception of morality is that morally good conduct is conduct affecting others that is in accord with values, dispositions, and character traits the cultivation of which increases one's prospects for long-term happiness. The empirical basis of the conception is a view in which powerful social impulses relating us to others are part of our nature, and hence, for all practical purposes, inescapable. Since they are not the only determinants of our behavior, we do not always act in accordance with them. Since they are so important to our well-being, however, we always have a reason to honor them, and we risk violating them at our peril. For Schlick, this is the true morality in which virtue and happiness are complementary, rather than at odds with one another.

He is convinced that this account of human psychology and human motivation is correct. The only thing that gives him pause is, I think, the metaethical claim that the conception of morally good conduct we have sketched is what the expression 'morally good conduct' literally means.

[21] Ibid., pp. 196–97.

He says, in the passage cited, that "the philosopher could, for his purposes, *define* as moral" the behavior we outlined. He adds the following:

> We must not forget, however, that he [the philosopher who does so define 'moral'] would in this fashion establish nothing but a definition, at bottom arbitrary, as is every other. He cannot force one to accept it, and cannot elevate it into a "postulate." I would hold it practical to accept this definition, because the end it establishes is that which *de facto* is most highly valued by mankind.[22]

I can't help thinking this disclaimer is a bit confused. If every definition is arbitrary, then the definition of a *square* as a rectangle with equal sides is arbitrary. But that doesn't mean it isn't a correct definition of 'square'. So why should we assume that the imagined "philosopher's definition" of 'moral' isn't correct? Of course, one can't *force* anyone to accept it, or any other definition. But often one can find linguistic evidence supporting the claim that a definition correctly gives the meaning of a word—in which case the definition isn't a postulate either. Is it obvious that no such empirical evidence could be found for the "philosopher's definition"? I doubt Schlick was of one mind about this. What, after all, is the point of his "practical" suggestion that we *accept* the definition because the coincidence of virtue and happiness it establishes is so highly valued? How could that be a reason for accepting the definition if we didn't already believe—prior to philosophical argument—that the very fact that we expect the claims of morality to motivate us shows that apparent conflicts between virtue and happiness must, in the end, be resolvable? Finally, "the philosopher's" claim about the connection between moral goodness and happiness doesn't have to be seen as a *definition* at all. Instead, Schlick could treat it as an empirical generalization about the connection between virtue and happiness that is confirmed by verifiable truths about human psychology plus metaethical observations about the truth conditions ethical statements must have if they are to fulfill their linguistic functions.

Whatever qualms Schlick may have felt about advancing surprising and far-reaching normative claims of this sort that reflected his ethical beliefs are submerged in the last few pages of his book, where he appears to speak with full-throated moral authority. Here is a sample of his remarks.

> [I]t seems to me that the idea of the capacity for happiness must everywhere be made central in ethics. And if a moral principle is needed it can only be one which rests upon this concept, as does the formula just proposed [in the last cited passage]. Therefore it is truly amazing that readiness for happiness nowhere plays an important role in ethical systems [proposed by philosophers].[23]

[22] Ibid., p. 197.
[23] Ibid., p. 198.

> A necessary condition of the capacity for happiness is the existence of inclinations in which the motive pleasure and pleasure of realization do not clash; and all conduct and motives which strengthen such inclinations are to be accepted as leading to the most valuable life. Experience teaches that these conditions are fulfilled by the social impulses, hence by those inclinations which have as their goal the joyful states of other creatures; with them there is the least probability that these joys of realization do not correspond to the motive pleasure. They are, if we use the philosophical definition of morality . . . the moral impulses *par excellence*. I am, in fact, of the opinion that those philosophers are quite right . . . who find the essence of moral dispositions in *altruism*. We recognized that its essence lies in *considerateness* for one's fellow men; in accommodation to and friendly understanding of their needs lies the very essence of the moral character. . . . Considerateness consists in the constant restraint and restriction of the non-altruistic impulses; and one can perhaps conceive all civilization as the colossal process of this subjugation of egoism. [24]

> As soon as the altruistic and the . . . "higher" impulses [pleasure in knowledge, beauty, etc.] are developed to a sufficient degree the process of subjugation [of egoism] is completed . . . for proper conduct . . . now flows quite of itself from the harmonious nature of the man. He no longer falls into "temptation"; "moral struggles" no longer occur in him. . . . There is no longer required a strong excitation of pain to deter him from . . . [bad] ends. . . . This is, of course, wholly attained by no one. And thus civilization works ceaselessly with all its means to establish motives for altruistic conduct.[25]

4. MIGHT THERE BE A SCIENTIFIC BASIS FOR MORALITY?

The final view articulated in the last chapter of *Problems of Ethics* places Schlick in a tradition that attempts to ground morality in human nature, as Aristotle does, or in the *moral sentiments*, as David Hume and Adam Smith do. Though the tradition is distinguished, it did not win very many philosophical adherents in the twentieth century. However, it did win some adherents outside of philosophy, including the celebrated social scientist John Q. Wilson, whose 1993 volume *The Moral Sense* sought to make something like the moral-sense tradition scientifically respectable. The author of ground-breaking work on the causes and prevention of crime, Wilson became fascinated by the norm from which criminals deviate. What leads so many people to behave in ways that are mostly lawful, moral, and other-regarding, even in the absence of threats or coercion? The answer,

[24] Ibid., pp. 199–200.
[25] Ibid., pp. 200–202.

he suggests, is an inherent moral sense consisting of a complex of social dispositions relating us to our fellows. If he is right, our genetic endowment, our early family experience, and the unalterable circumstances of the human condition provide us with a motivational base that ties us by bonds of affection, social affiliation, and mutual interest to our fellows in ways that Schlick would recognize. This, they both suggest, is the raw material that generates reasons for other-regarding action, the authority of which can be recognized by most human beings.

In making his case, Wilson repudiates Freud and embraces Darwin. Because cooperation promotes survival, we have, he argues, been bred by natural selection to be social animals. It is not just that we need and want what others can provide, and so are impelled by self-interest to depend on them. We are also disposed to form powerful cognitive and emotional *attachments* to them. Parents are innately disposed to protect, nurture, and love their young. Children naturally bond with parents, while emulating not only their parents, but also others with whom they are intimate. In their early years they form reciprocal bonds of affection and trust in which their well-being and self-conception becomes intertwined with others. Entering into games and collective activities, they learn the rudiments of fairness, which involves adhering to common rules and earning rewards proportional to the value of their efforts.

According to Wilson, this fusion of natural sentiment with rational principle gives birth to morality. Sentiment infuses our participation in games and collective activities with those we like and admire, and who we hope will like and admire us. Often these companions will be models of the people we wish to become. The rules governing our activities with them are often impersonal principles that apply to anyone who occupies a given role in the effort. Because these rules define the commonly accepted terms of participation in a mutually beneficial undertaking, it is in the self-interest of each participant to obey them. But they are more than prudential rules of thumb. Because the parties are often comrades bound by ties of social affiliation, rule violations carry psychic risks beyond the loss of the purely self-regarding benefits secured by participation. Violations of rules governing interaction with one's socially affiliated fellows are *affronts* to one's comrades, to one's friendship with them, to one's image in their eyes, and to the person one wants, with their help, to become. With this, instrumentally useful rules obeyed to secure the benefits of group action become principles to be honored even when no one is looking. This is the point at which sentiment, social affiliation, and recognition of mutual interest are incorporated into the binding commitments and broad principles that are the foundation of morality.

This Schlick-friendly sketch of how Wilson's moral sense may generate embryonic moral principles defining obligations to family, friends, and compatriots is, of course, only the beginning. Much more is needed to explain how broader commitments might be generated—to casual

acquaintances, to strangers one encounters, to one's community, to one's country, and even to all human beings. Though this explanatory task is far from complete, and is to some degree speculative, it suggests that the brand of ethical naturalism espoused by Schlick is not a mere historical curiosity, but continues to be pursued in sophisticated ways.

CHAPTER 13

Replacing Ethics with Metaethics: Emotivism and Its Critics

1. EMOTIVIST DOCTRINES AND THE ARGUMENTS FOR THEM

The emotivist theory of value is a well-known and influential philosophical view which, although an important part of logical empiricism, is conceptually detachable from it. It was part of logical empiricism because several of its main tenets appeared to be supported by the verifiability criterion of meaning. It is detachable from logical empiricism because it also had other sources of support. Consequently, it was able to survive and evolve into other forms of non-cognitivism after classical verificationism had fallen by the wayside. Two leading emotivists I will consider are A. J. Ayer, who presented his views in chapter 6 of *Language, Truth, and Logic*, and Charles L. Stevenson, whose views are presented in his seminal paper, "The Emotive Meaning of Ethical Terms."[1]

I begin with four central claims made by Ayer.[2]

[1] Stevenson (1937 [1959]).

[2] Ayer and other logical empiricists didn't carefully distinguish sentences, statements, and judgments. In order to achieve reasonable fidelity to the texts, I will, for the most part, follow this regrettable tendency in reporting their views, except where more precise reformulation is required.

E1. No evaluative judgment (sentence/statement) of the form 'x is good/bad/right/wrong/morally required is equivalent to any nonevaluative judgment (sentence/ statement).[3]
E2. No nonevaluative judgment (sentence/statement) entails any evaluative judgment (sentence/statement) of the form 'x is good/bad/right/wrong/morally required.[4]
E3. No evaluative judgment (sentence/statement) of the form 'x is good/bad/right/wrong/morally required entails any nonevaluative judgment (sentence/statement).
E4. Evaluative judgments (sentences/statements) are neither true nor false. They do not state facts. Rather, their meaning is entirely emotive.

There were three main argumentative routes to these theses. The first was verificationist. Since evaluative sentences seem not to be used to make statements that are verifiable by empirical observation, emotivists took them to be cognitively meaningless. Thus, at the very beginning of his discussion of ethics, Ayer takes it to be beyond dispute that uses of ethical sentences "cannot with any show of justice be represented as [expressing genuine] hypotheses, which are used to predict the course of our sensations."[5] Responding to the idea that uses of ethical sentences make statements that can be known to be true, or false, by "intellectual intuition," he replies that empirical meaningfulness requires observation-based criteria to resolve conflicts among such "intuitions." These, he assumes, do not exist.[6] Hence, he concludes that ethical sentences are incapable, when used in accordance with the linguistic conventions that govern them, of expressing statements that are either true or false; if they are to have any function at all, their function must be non-cognitive, or emotional.

Carnap reasons in the same way in *Philosophy and Logical Syntax*.

> [A value statement] does not assert anything and can neither be proved nor disproved. This is revealed as soon as we apply to such statements our method of logical analysis. From the statement "Killing is evil" we cannot deduce any proposition about future experiences. Thus this statement is not verifiable and has no theoretical sense, and the same thing is true of all other value statements. . . . The propositions of normative ethics, whether they have the form of rules or the form of value statements . . . are not scientific propositions (taking the word scientific to mean any assertive proposition).[7]

The second argumentative route to emotivism began with G. E. Moore. The emotivists accepted Moore's critique of ethical naturalism. According

[3] Ayer (1936 [1946]), pp. 104–5.
[4] Ibid., pp. 104–5.
[5] Ibid., p. 102.
[6] Ibid., pp. 104–5.
[7] Carnap (1935a), pp. 24–25. The book is derived from three lectures Carnap gave at the University of London in October 1934.

to Moore, the central evaluative notion, 'good', is indefinable, and so can't stand for a complex property. Nor, he thought, could it stand for any simple natural property the presence or absence of which can be settled by observation. Accepting Moore's argument, emotivists agreed that 'good' doesn't stand for any natural property. Here is how Stevenson puts it.

> The omnipotence of [a certain kind of descriptivist theory of the meaning of 'good'] . . . may be shown to be unacceptable in a somewhat different way. Mr. G. E. Moore's familiar objection about the open question is chiefly pertinent in this regard. No matter what set of scientifically knowable properties a thing may have (says Moore, in effect), you will find, on careful introspection, that it is an open question to ask whether anything having these properties is *good*. . . . [So] we must be using some sense of "good" which is not definable, relevantly, in terms of anything scientifically knowable.[8]

The emotivists then went beyond Moore. Whereas he concluded that 'good' must stand for a non-natural property, they rejected non-natural properties as mysterious we-know-not-whats, and concluded that 'good' doesn't express any property at all. Rather, they thought, its function is to express emotions.

One possibility not taken seriously was that ethical primitives, like theoretical primitives in science, express indefinable natural properties we gain knowledge of empirically—perhaps, in the case of ethics, by reflecting on our own experience and studying human nature. Even the prominent critic of emotivism Sir David Ross seems to have taken it for granted that this was not a coherent option. Writing in *The Foundations of Ethics,* he says:

> If Subjectivism and Utilitarianism are rejected, as they are by the positivists, it might seem that the conclusion to be drawn is that 'right' and 'good', and their opposites, are terms which cannot be defined naturalistically, and that judgments in which we use them as predicates are *a priori* judgments, judgments in which we express not the results of observation, but a direct insight.[9]

By "judgments in which we use ethical terms as predicates," Ross clearly does not mean judgments that particular individuals are good/bad or that specific acts they perform are right/wrong. Rather, he must be thinking of general moral principles. He simply assumes that if there are any general ethical truths, they must be synthetic, necessary a priori principles known by "direct insight" or intuition. Unfortunately, this unilluminating moral epistemology was widely seen as the only serious cognitivist alternative to emotivism, lending it strength it might otherwise not have had. It is a pity, in this respect, that Schlick didn't do a better job in articulating his empirical conception of ethics.

[8] Stevenson (1937 [1959]), p. 268.
[9] Ross (1939), pp. 32, my emphasis. The book is derived from Ross's 1935–36 Gifford Lectures delivered at the University of Aberdeen.

The third argumentative route to emotivism rested on the action-guiding character of evaluative language, which was emphasized by Stevenson. Emotivists thought that to sincerely say of something "It's good" is to have a positive emotional attitude toward it. To sincerely say of a prospective action "I ought to do it" is to express a positive motivation for performing it. To say "You ought to do it" or "It's right" about an act another might perform is to urge its performance. For emotivists, it is part of what we mean by words like 'good' and 'right' that anyone who, at the time of utterance, has no positive feelings toward x *cannot* sincerely say of x "It's good" or "It's right." In short, they held that *if one sincerely characterizes x to be good or right (at t), then x cannot leave one cold (at t)*.

Emotivists used this principle against theories of the meanings of evaluative words like 'good' and 'right' according to which they are used to state genuine facts about the properties of individuals, actions, and other things in the world. Thus, Stevenson says:

> "goodness" must have, so to speak, a magnetism. A person who recognizes X to be "good" must *ipso facto* acquire a stronger tendency to act in its favor than he otherwise would have had. This rules out the Humian type of definition. For according to Hume, to recognize that something is "good" is simply to recognize that the majority approve of it. Clearly, a man may see that the majority approve of X without having, himself, a stronger tendency to favor it. This requirement excludes any attempt to define "good" in terms of the interest of people *other* than the speaker.[10]

The general point is that it is possible for a person to sincerely judge an action to be one that (i) produces the greatest happiness for the greatest number, (ii) promotes human survival, (iii) is approved of by most people, or (iv) is what God wants us to perform, without having any positive feelings about the action, or recognizing any motivation to perform it. Since emotivists took it to be impossible for a person to sincerely characterize an action as "good" or "right" without having such feelings and recognizing such motivations, they concluded that 'good' and 'right' can't mean the same as *action that produces the greatest happiness for the greatest number, action that promotes human survival*, *action approved of by most people,* or *action that God wants us to perform*.

At most this argument shows that *good* and *right* are not strictly synonymous with any descriptive phrase *D* the meanings of which don't encode any intrinsic connection to motivation. In itself, this result is rather weak, and analogous to Moore's conclusion that goodness cannot be descriptively defined. To turn the emotivists' conclusion into something stronger, one needs two things: (i) a rejection of the synthetic, necessary a posteriori (which Moore accepted) plus (ii) the bundle of Moore's flawed

[10] Ibid., pp. 266–67.

assumptions about *synonymy, definition, analyticity, logical consequence,* and *entailment*.[11] Although (i) and (ii) were congenial to logical empiricists—with their one-dimensional conception of the modalities—they ceased to be available to non-cognitivists when the link between non-cognitivism and its logical empiricist origins was severed.

2. EMOTIVISM, METAETHICAL EGOISM, AND ETHICAL DISAGREEMENTS

Despite the apparent power of the emotivists' action-guiding principle, there was one descriptivist theory of the meanings of evaluative terms—metaethical egoism—that emotivists like Stevenson recognized to be consistent with it. According to this theory, the sentence 'Telling the truth is right' is used to state that the speaker prefers people to tell the truth. Since it would be incoherent to sincerely assert that one prefers people to tell the truth and then go on to add *but telling the truth leaves me cold; I am completely indifferent to it*, metaethical egoism is compatible with the emotivists' observation that evaluative judgments are emotive and action-guiding.

But it isn't compatible with emotivism. According to egoism, evaluative sentences are used to make true or false statements about what one prefers. This conflicts with the emotivist doctrine that evaluative sentences are not used to state facts, or indeed to make any statements. Since the emotivists' action-guiding principle was powerless against metaethical egoism, they needed another argument against it.

Stevenson (1937) provided it by resurrecting Moore's old argument about disagreement.[12] Imagine the following dialog between A and B.

A: Fighting terrorists is the right thing to do.
B: That's not so. Fighting terrorists is not right. We should try to understand them.

According to egoism this dialog is equivalent to:

A'. I prefer that we fight terrorists.
B'. That's not so. I prefer that we not fight terrorists. We should try to understand them.

This egoist analysis of the dialog misconstrues B's response to A. In saying "That's not so," B *contradicts* A. The egoist's analysis misses this by interpreting A and B simply as making compatible statements about their own preferences.[13]

[11] See Soames (2014), chapter 4.
[12] Moore's argument is discussed in Soames (2014), chapter 6.
[13] Stevenson notes this on p. 266 of Stevenson (1937 [1959]). He says: "[W]e must be able to sensibly *disagree* about whether something is 'good.' This condition rules out Hobbes's

It is hard not to agree with Moore and Stevenson that the egoist's analysis of the disagreement between A and B is unacceptable.[14] But can the emotivist provide a better one? If emotivism merely notes that the speakers express different emotions, it won't explain the *disagreement* between A and B. To give an explanation, one must recognize more than simple displays of raw emotion. Recognizing this, Stevenson maintained that many uses of evaluative language can be analyzed as making recommendations, rather than as making statements that are true or false. On his view, the dialog between A and B can analyzed along the following lines:

A*. Let's all support the fight against the terrorists.
B*. On the contrary, let's not support the fight against them. Instead, let's try to understand them.

Here, A and B are seen as making conflicting recommendations (that can't jointly be followed) rather than conflicting statements (that can't jointly be true). Stevenson labels this a *disagreement in interest* rather than a *disagreement in belief*.

How are disagreements in interest to be resolved? According to emotivists, many arise because people have different factual beliefs. The way to resolve these evaluative disagreements is to achieve agreement on the relevant nonevaluative, empirical facts. Here is how Ayer puts it.

> When someone disagrees with us about the moral value of a certain action or type of action, we do admittedly resort to argument in order to win him over to our way of thinking. But we do not attempt to show by our arguments that he has the "wrong" ethical feeling towards a situation whose nature he has correctly apprehended. What we attempt to show is that he is mistaken about the facts of the case. We argue that he has misconceived the agent's motive; or that he has misjudged the effects of the action, or its probable effects in view of the agent's knowledge. Or that he has failed to take into account the special circumstances in which the agent was placed.[15]

Ayer here focuses on cases in which disagreement about the moral assessment of an act performed or contemplated is based on empirical disagreements about the motive, its consequences, or the special circumstances in which the agent is placed. In such cases the disagreements arise from

---

[egoistic] definition. For consider the following argument: 'This is good.' 'That isn't so. It's not good.' As translated by Hobbes, this becomes 'I desire this.' 'That's not so, for I don't.' The speakers are not contradicting one another. . . . The definition, 'good' means *desired by my community*, is also excluded, for how could people from different communities disagree?"

[14] Ayer (1936 [1946]) takes a slightly different view of Moore's argument on p. 110. He builds into Moore's argument not only the assumption that there is a genuine disagreement of some sort between A and B, but also the assumption that the disagreement is a factual one. Thus, he takes Moore's argument to incorporate a false assumption. I separate the two assumptions, and take the argument to incorporate only the first.

[15] Ayer (1936 [1946]), pp. 110–11.

different factual beliefs. For example, the disagreement between A and B might be caused by an underlying factual disagreement about (i) the causes that led the terrorists to perpetrate their attacks, (ii) their ultimate goals and motivations, (iii) the prospects for, and costs of, defeating them and their allies militarily, (iv) the likelihood that future terrorism can be deterred by swift and strong military action, and (v) the likelihood that restraining the military and compromising with the terrorists would encourage others to launch similarly violent attacks in the future. All of these are straightforward empirical matters that could, in principle, be rationally investigated. Emotivism can accommodate this.

But this appeal to rationality will work only when the disagreements are based solely on different beliefs about a nonevaluative matter. If A and B have different fundamental values—different basic preferences about certain kinds of conduct, various forms of social organization, or personal interaction, or about other fundamental matters—then emotivism maintains that there can be *no* rational resolution of their differences. Consider an example. Suppose that A values punishing and even putting to death those who have murdered thousands of innocent people, not simply because doing so will deter others, but also because our sense of justice demands it. Suppose, on the other hand, that B abhors revenge and violence in any form, and would not favor retributive violence or capital punishment under any circumstances. If these different attitudes of A and B are *not* based on different factual beliefs, then emotivism tells us that a rational resolution of the evaluative differences between them is *conceptually impossible*.

It is important to realize that the emotivists were not making a psychological or sociological point. It is not just that A and B might never come to agree about capital punishment. Nor is it that, human nature being what it is, one can't expect people to be rational about things they hold dear. The emotivists' point is that when A says, in the situation described, "Capital punishment of mass murderers is right" and B says "It's wrong," there is no factual issue, there is no genuine belief whatsoever, separating them. Since there is no belief on which they differ, *there is nothing separating them about which it is possible to reason*. Their difference is entirely a difference in interest.

3. CRITICISMS OF EMOTIVISM

3.1. Cognitive Disagreement in Ethics

According to Ayer, when ethical disagreement about an action has exhausted all possible sources of purely factual disagreement over motives, consequences, and special circumstances, there can be no rational inquiry into, or argumentative resolution of, purely evaluative differences. He signals his commitment to this view in the last cited passage by observing

that when empirical sources of disagreement are exhausted, we don't try to show that our interlocutor has "the wrong ethical feeling." The remark is problematic, because no one would think otherwise. Non-emotivists don't think that what ordinarily passes for different *beliefs* about the right and the good are merely different *feelings*. They don't criticize their opponents' feelings; they criticize them for missing important ethical truths. Emotivists don't criticize feelings either, because they take feelings unconnected with cognition not to be rationally criticizable.

Ayer fills out his discussion of the limits of ethical argumentation by completing the above passage as follows:

> We do this [focus on empirical differences] in the hope that we have only to get our opponent to agree with us about the nature of empirical facts for him to adopt the same moral attitude towards them as we do. And as the people with whom we argue have generally received the same moral education as ourselves, and live in the same social order, our expectation is usually justified. But if our opponent happens to have undergone a different process of moral "conditioning" from ourselves, so that, even when he acknowledges all the facts, he still disagrees with us about the moral value of the actions under discussion, then we abandon the attempt to convince him by argument. We say that it is impossible to argue with him because he has a distorted or undeveloped moral sense; which signifies merely that he employs a different set of values from our own. We feel that our own system of values is superior. . . . But we cannot bring forward any arguments to show that our system is superior. For our judgment that it is so is itself a judgment of value, and accordingly outside the scope of argument.[16]

Though not without merit, this conception of moral argument is one-sided. Yes, sustained moral inquiry must, if it is to lead us to reasons for acting morally, ground those reasons in values we are capable of recognizing as our own. But recognizing which of our values bear on an action in which ways is a complicated matter that can't be equated with simply registering spontaneous feelings that arise when we contemplate the action, even if we have accurately assessed its consequences and the context in which it occurs. Our fundamental values concerning others are complex and intricately interconnected. All of us want and need emotional attachments to others, whose help, affection, and good will demand similar help, affection, and good will from us. All of us internalize standards of conduct for others that we wish to live up to ourselves, and to be seen as living up to. We also value fair processes for participating in joint actions and equitably sharing in its risks and rewards. Recognizing our own mortality, we wish to be honored and remembered, and we have a powerful interest in identifying with human projects and groups larger than ourselves.

[16] Ibid., p. 111.

Ayer's passage doesn't mention these fundamental human interests, but focuses instead on similarities in moral education and social background as sources of common values. These are not irrelevant, but they are not as morally significant as deeper values that are much more widely shared. He also misses a related aspect of the reality of ethical inquiry, argument, and debate. Because we have many social values that are capable of interacting in unanticipated ways in new and unusual cases, we don't always know which actions best advance our values. In these cases, we may turn to reasoning about counterfactual circumstances to sharpen and resolve evaluative conflicts. For example, in both traditional philosophical discussions and ordinary life, we may engage in thought experiments to sharpen and resolve conflicts of liberty versus equality in our social lives, of outcomes based on fair procedures versus outcomes that maximize general welfare, of injunctions to benefit others versus injunctions not to harm innocent third parties, and of fidelity to promises and other voluntarily undertaken obligations versus the benefits for oneself and others that can be secured by violating them. In none of these cases is our inquiry, debate, and argument restricted to nonevaluative empirical questions about what exists now, to what has occurred in the past, or even to what will occur in the future.

Often the inquiry involves consideration of counterfactuals (which were not well understood by logical empiricists) about what would occur if certain scenarios were realized. Typically, this will set the stage for moral evaluations of actions, events, and states of affairs in those scenarios. On the basis of these evaluations, agents revise or extend the moral principles they have heretofore explicitly recognized, thereby modifying their moral commitments. For example, I may come to think that the sacrifices in liberty inherent in John Rawls's *difference principle*—which asserts that justice requires a social distribution of material goods that maximizes the welfare of the least well off—are too high to be morally acceptable. This in turn may lead me to reverse my previous evaluative judgments of some real-world actions and policies, and to form new moral judgments in other cases about which I now hold no determinate view.

Consider a case of "disagreement" of this sort between my former self and my present self. Prior to my counterfactual reasoning, I said of a certain action "It's right." Afterwards, I say "I was wrong; the action isn't right." Let us stipulate that this change in view was not accompanied by any change in my beliefs about actual (as opposed to counterfactual) nonevaluative factual matters, past, present, or future. What, I wonder, would Ayer identify as the target of my criticism of my past self? What, from my present perspective, was I wrong about? Not, as he insists in the earlier passage, my past *feelings*. I don't criticize those feelings as wrong. Nor, as I have stipulated, do I reject any nonevaluative belief I had about a past, present, or future matter of actual fact. Ayer could observe that, as a result of my counterfactual reasoning, I came to have new beliefs about the

extent of the restrictions on liberty that *would occur if the action in question, or those like it, were performed*. He could also correctly point out that these should count as new empirical beliefs. But this wouldn't show that I was *wrong* about anything, if, as might well have been the case, I simply had not previously considered the complex interaction between liberty and the limited Rawlsian version of equality. In such a case, the only *past* beliefs I now maintain to have been in error were my moral beliefs. Indeed, I could have explicitly indicated this by saying of the action in question, "Although I incorrectly *believed* otherwise, I now see that the action isn't right." Ayer can't admit this, because he denies that there is such a thing as a moral belief.

His problem stems, at least in part, from too narrow an understanding of the magnetic character of moral uses of words like 'good', 'right', and 'morally required'. Uses of these words are connected with actions, which, in turn, are guided by values. But the connection that is crucial to sincere characterizations of an act as "good," "right," or "morally required" is not a direct connection to the *feelings* of the agent at a given time; it is a connection to the interests or values of the agent that is mediated by agent's moral *judgments*.

With this in mind, consider the following anti-emotivist view.

INTERNALIST MORAL REALISM

(i) An agent A who sincerely says of an act A is contemplating performing "It is morally required" expresses the belief that A is morally required to perform it.
(ii) A is morally required to perform an act only if A has a reason to perform it.
(iii) A has a reason to perform an act only if it advances A's fundamental interests or values.

Unlike emotivism, this view is cognitivist. It holds that uses of moral sentences make statements that are true or false. Nevertheless, if (ii) is widely recognized to be true, judging that an act one contemplates performing is morally required will typically involve *believing that one has a reason to perform it*. The suggested cognitivist view of moral language and moral obligation will, therefore, agree with emotivism in one respect: because acts one characterizes as "morally required" are those one has some inclination to perform, they are not acts that *leave one cold*. Finally, if (iii) is true, then one's belief that one is morally required to perform a given act can be false, because, despite what one may have thought, it doesn't advance one's fundamental values. One can be in this state of evaluative error when one is wrong about what fidelity to one's own values demands—as I take myself to have been prior to my thought experiment involving the interaction of liberty with Rawlsian equality.

Of course, the view just sketched is not a fully fledged ethical or metaethical theory. Presumably, many moral claims made by an agent don't *state* anything about the agent or the agent's own values. In addition, no matter what the content of these statements turns out to be, their truth or

falsity must depend on more than the individual agent's fundamental motivational set, and the reasons for action it provides. The view also requires a robust conception, along the lines hinted at in chapter 12, of the biologically determined social impulses of normal human beings that link us to others and that intertwine our welfare with theirs. The point here is that it *might* be possible to construct an ethical and metaethical view of this sort that stands some chance of success. Such a view is intriguing and worth pursuing, in part because of its conceptual links both to Schlick's empirical cognitivist conception of ethics and to the non-cognitivist conception of his emotivist colleagues.

3.2. The Problem of Evaluative Entailments

The next problem grows out of attempts to generalize Ayer's theses E1–E3.

E1. No evaluative judgment (sentence/statement) of the form 'x is good/bad/right/wrong/morally required' is equivalent to any nonevaluative judgment (sentence/ statement).
E2. No nonevaluative judgment (sentence/statement) entails any evaluative judgment (sentence/statement) of the form 'x is good/bad/right/wrong/morally required'.
E3. No evaluative judgment (sentence/statement) of the form 'x is good/bad/right/wrong/morally required' entails any nonevaluative judgment (sentence/statement).

These theses presuppose an exhaustive dichotomy: every sentence/statement is either evaluative or nonevaluative. Presumably sentences that don't contain any evaluative words are nonevaluative. It would be natural to suppose that sentences that do contain evaluative words count as evaluative sentences. With this understanding, we might generalize E1–E3 as follows. Examples (1)–(3) show that this is problematic.

E1*. No evaluative judgment (sentence/statement) is equivalent to any nonevaluative judgment (sentence/statement).
E2*. No nonevaluative judgment (sentence/statement) entails any evaluative judgment (sentence/statement).
E3*. No evaluative judgment (sentence/statement) entails any nonevaluative judgment (sentence/statement).

Now there is trouble.

1. You stole that money.
2. You acted wrongly in stealing that money.
3. Stealing money is wrong.

Here (3) is clearly evaluative and (1) is apparently nonevaluative. (If 'stole' is already a moral term, change the example.) But what about (2)? Since it contains an evaluative word, it would seem to be evaluative. But then, since (2) entails (1), which is nonevaluative, we falsify E3*.

Ayer discusses these examples in *Language, Truth, and Logic*.

> The presence of an ethical symbol in a proposition adds nothing to its factual content. Thus if I say to someone, "You acted wrongly in stealing that money," I am not stating anything more than if I had simply said, "You stole that money." In adding that this action is wrong I am not making any further statement about it. I am simply evincing my moral disapproval of it. It is as if I had said it with the addition of some special exclamation marks. The tone, or the exclamation marks, adds nothing to the literal meaning of the sentence. It merely serves to show that the expression of it is attended by certain feelings in the speaker.[17]

He seems to say that sentences (1) and (2) are used to make the same statement, and so have the same literal content or meaning. If so, they must be logically equivalent, in which case they contradict E1*.

A natural response might hold that (2) is a complex sentence including an *evaluative* part and an *empirical* part. On this view, the logical form of (2) is something like

2'. You stole that money and stealing money is wrong (or that was wrong).

The left-hand conjunct is an *empirical* sentence and the right-hand conjunct is a purely *evaluative* sentence. What about the compound sentence? We can't call it *evaluative*, because it entails the left-hand conjunct, which is empirical. We can't call it *empirical*, because it entails the right-hand conjunct, which is evaluative. We could say that (2) and (2') are *mixed sentences*, thereby recognizing three kinds of sentence.

We could then restate theses E1*–E3* as follows:

E1**. No evaluative judgment (sentence/statement) is logically equivalent to any empirical judgment (sentence/statement).
E2**. No empirical judgment (sentence/statement) entails any evaluative judgment (sentence/statement).
E3**. No evaluative judgment (sentence/statement) entails any empirical judgment (sentence/statement).

In addition to advancing these theses, the emotivist will continue to hold that evaluative sentences do not state fact, are neither true nor false, and have meanings that are entirely emotive. But, although this is progress, there is still a problem. According to standard definitions of entailment, it is a species of truth preservation. Hence, things that stand in this relation to one another are things that are capable of being true or false—statements, propositions, or sentences used to make statements, or express propositions. We don't say that a cheer, a grunt, a smile, an exclamation—*Wow!*—or even a command *entails* any statement that is true or false.

[17] Ibid., p. 107.

With this in mind, consider again the observation that (2) entails (1). This might seem to be all right, in view of the fact that (2) is a mixed sentence, having the logical form (2'). But how is (2') to be understood? Since its right-hand conjunct is evaluative, it would seem that, according to at least some prominent emotivist analyses, (2') ought to be understood along the lines of (2").

2". You stole that money and don't steal money!

But does it really make sense to say that (2"), as a whole, is the sort of thing that can be true or false? If not, then the emotivist cannot contend that (2) has a truth value, and so cannot admit what seems to be an obvious fact—namely, that (2) does entail (1).

We can put the worry more strongly. (3) is a *purely* evaluative sentence, which, according to the emotivist, ought to have a logical form along the lines of (3').

3'. Don't steal money!

But it seems clear, pretheoretically, that (3) entails the conditional the antecedent of which is sentence (1), and the consequent of which is sentence (2). In other words, (3), 'Stealing money is wrong', entails (4).

4. If you stole that money, then you acted wrongly in stealing that money.

But whatever one says about (2) being true or false, the emotivist *must* claim that (3) is incapable of being true or false. Thus, it is not clear that the emotivist can capture our pretheoretic conviction that (3) enters into genuine entailment relations.

This is a serious problem for classical emotivism, but perhaps not an insoluble one. Evaluative sentences do enter into conceptual relations of some sort with various types of sentences. Emotivists can't explain this by appealing to the standard notion of logical entailment. Hence, they must attempt to explain it some other way. To do this, they first need to be more specific about how evaluative uses of language are to be understood. Are they exclamations, are they equivalent to utterances of imperatives, are they performances of some sort—e.g., commands or recommendations? Next, emotivists need to characterize conceptual relations different from, but analogous to, logical entailment that evaluative sentences can bear not only to one another but also to nonevaluative sentences. Steps taken by descendants of the original emotivists to address this problem will be discussed in later volumes.

3.3. The Emotivists' Performative Fallacy

Emotivism was put forward as a theory of the meaning of evaluative words like 'good', 'bad', 'right', 'wrong', 'just', 'unjust', 'should', 'ought', and so on. The theory attempted to specify the meanings of these terms by

specifying the meanings of simple sentences used with the intention of *calling* something good, bad, right, wrong, just, unjust, and the like—sentences like those in (5).

5a. That book is good.
b. Stealing is wrong.
c. The government is unjust.

We have seen that Ayer and Stevenson analyzed the meanings of these sentences in terms of the kinds of linguistic acts—such as giving commands, issuing orders, and making recommendations—that speakers perform when they uttered them. Carnap agreed.

> It is easy to see that it is merely a difference of formulation, whether we state a norm or a value statement. A norm or rule has an imperative form, for instance: "Do not kill!" The corresponding value judgment would be "Killing is evil." The difference of formulation has become practically very important, especially for the development of philosophical thinking. The rule, "Do not kill," has grammatically the imperative form and will therefore not be regarded as an assertion. But the value statement, "Killing is evil," although, like the rule, it is merely an expression of a certain wish, has the grammatical form of an assertive proposition. Most philosophers have been deceived by this form into thinking that a value statement is really an assertive proposition, and must be either true or false. Therefore they give reasons for their own value statements and try to disprove those of their opponents. But actually a value statement is nothing else than a command in a misleading grammatical form.[18]

Because of this "performative analysis" of uses of ethical sentences, emotivists imagined that the meaning of 'Stealing is wrong' or 'One ought not to steal' was roughly the same as the meaning of 'Don't steal!' In each case, it was assumed that to know the meaning of the sentence is to know that uses of it are orders or commands that one not steal. Similarly, the meaning of 'That is good' was thought to be something like the meaning of 'I recommend x'. It was assumed that to utter this sentence is *not* to assert that one is recommending x, but to perform the act of recommending itself.

That, very roughly, was the structure of the emotivist view, which suffered from a fatal flaw. If one hopes to give a theory of meaning of evaluative words, phrases, and sentences, one can't restrict oneself to a limited range of linguistic environments. In particular, one can't restrict oneself to simple sentences utterances of which are acts of recommending, commanding, and the like. Rather, one's semantic theory of evaluative expressions must apply to all sentences in which they occur.

The first critic of emotivism to make this point seems to have been Sir David Ross.

[18] Carnap (1935a), pp. 23–24.

> The theory that all judgments with the predicate 'right' or 'good' are commands has evidently very little plausibility. The only moral judgments of which it could with any plausibility be maintained that they are commands are those in which one person says to another 'you ought to do so-and-so'. A command is an attempt to induce some one to behave as one wishes him to behave, either by the mere use of authoritative or vehement language, or by this coupled with the intimation that disobedience will be punished. And there is no doubt that such words as 'you ought to do so-and-so' may be used as one's means of so inducing a person to behave a certain way. *But if we are to do justice to the meaning of 'right' or 'ought', we must take account also of such modes of speech as 'he ought to do so-and-so', 'you ought to have done so-and-so', 'if this and that had been the case, you ought to have done so-and-so', 'if this and that were the case, you ought to do so-and-so', 'I ought to do so-and-so'. Where the judgment of obligation has reference either to a third person, not the person addressed, or to the past, or to an unfulfilled past condition, or to a future treated as merely possible, or to the speaker himself, there is no plausibility in describing the judgment as a command.* But it is easy to see that 'ought' means the same in all these cases, and that if in some of them it does not express a command, it does not do so in any. *And if the form of words 'you ought to do so-and-so' may be used as a way of inducing the person addressed to behave in a particular way, that does not in the least imply that the apparent statement is not really a statement, but a command. What distinguishes its meaning from that of the genuine 'do so-and-so' is that one is suggesting to the person addressed a reason for doing so-and-so, viz., that it is right.* The attempt to induce the person addressed to behave in a particular way is a separable accompaniment of the thought that the act is right, and cannot for a moment be accepted as the meaning of the words 'you ought to do so-and-so'.[19]

The following criticism of emotivism is an elaboration of this line of argument.[20]

We begin by recognizing that although simple sentences containing evaluative words are often used to perform acts of recommending, commanding, and the like, they also occur in sentences like (6) that are not used in this way.

6a. George Bush Sr. should have finished off Saddam Hussein in 1991.
 b. I wonder whether I ought to work harder.
 c. If western-style democracies are just, then they will win the allegiance of their citizens.
 d. Bill hopes that electric blanket is a good one.

[19] Ross (1939), pp. 33–34, my emphasis.

[20] For some reason, Ross's original objection seems not to have attracted much attention. Decades later, Geach (1960) revived and elaborated the objection (without making reference to Ross). Geach's argument was directed at proponents of the ordinary language school at Oxford in the 1950s. The objection was elaborated further, without reference to Geach or Ross, in John Searle (1962). Searle does, however, cite the discussion of Paul Ziff (1960), section 227 and following, where a similar argument is developed.

It is hard to liken these sentences to imperatives, or to identify their meanings with any imaginable commands, orders, requests, or recommendations associated with them. Certainly, they can't be equated with the bizarre examples in (7).

7a. George Bush Sr., listen up, finish off Saddam Hussein in 1991!
George Bush Sr., I order you to finish off Saddam Hussein in 1991.
George Bush, Sr., I recommend that you finish off Saddam Hussein in 1991.
George Bush Sr., please, finish off Saddam Hussein in 1991.

7b. I wonder whether: work harder!
I wonder whether I order myself to work harder.
I wonder whether I recommend that I work harder.

7c. If: support western-style democracies!, then they will win the allegiance of their citizens.
If I order you to support western-style democracies, then they will win the allegiance of their citizens.
If I recommend western-style democracies, then they will win the allegiance of the citizens.

7d. Bill hopes: I recommend that electric blanket!
Bill hopes that I recommend that electric blanket.
Bill hopes that he recommends that electric blanket.
Bill hopes: buy that electric blanket, if you are in the market for one!

The lesson extends beyond amusing examples like these. Evaluative expressions occur in a wide variety of sentences; indeed, they occur in just about any linguistic environment in which arbitrary declarative sentences can occur. Hence, any theory of what evaluative expressions mean must explain their contributions to the meanings of sentences across the board. Emotivists failed to recognize this.

As a result, they missed the meanings of evaluative expressions when they occur in sentences like (6), which go beyond the restricted range of cases in which speakers use evaluative sentences to make straightforward recommendations, or to issue clear commands or orders. This failure supports the stronger conclusion that emotivism also fails to correctly characterize the meanings of the simple evaluative sentences, like those in (5), on which the emotivists concentrated so much of their attention. After all, it doesn't seem plausible that evaluative expressions *change* their meaning from one linguistic environment to another. When we consider the conditional sentence (6c), for example, it seems clear that it is intended to make a claim that can be evaluated for truth or falsity. But in order for this conditional to have a truth value, its *antecedent*, 'western-style democracies are just', must also have one—which means that it can neither be replaced by an imperative, nor be seen as here being used to make a recommendation. Presumably, we don't want to say that this simple sentence has a purely evaluative meaning—to be given solely in terms of imperatives or recommendations—when it is used all by itself, while having a different,

descriptive, meaning when it occurs as the antecedent of a conditional (or the complement of a propositional attitude verb like 'believe', 'hope', or 'wonder'). For if it did switch its meaning in this way, then the pattern of reasoning in (8) would be a simple piece of equivocation, rather than the deductively valid argument we recognize it to be.

8a. Western-style democracies are just.
b. If western-style democracies are just, they will win the support of their citizens.
c. Therefore, western-style democracies will win the support of their citizens.

In short, evaluative words *don't* have the meanings emotivism claimed them to have. This doesn't mean that they aren't used in simple sentences like those in (5) to make recommendations, to issue commands, or to exhort hearers to action. These sentences *are* used that way. But their *meanings* aren't given by specifying the evaluative acts they are often used to perform. Consider an analogous case. If I say, in a letter of recommendation, that a student is brilliant, I am thereby *praising* and *recommending* the student. These are linguistic actions. But the fact that I perform them doesn't show that the word 'brilliant' has a special, nondescriptive, performative meaning. It shows that to say that a student has a property, *being brilliant*, that we find desirable in students *is* to praise or recommend her.[21] By the same token, to say that something is good is often to recommend it, but this doesn't show that the word 'good' has a special performative meaning. It shows that the word 'recommend' is understood in such a way that one way to recommend something is to predicate goodness of it.

It is worth noting that some words—e.g., 'hello', 'ditto', 'please', and 'yes'—do have nondescriptive, performative meanings given by specifying the linguistic acts they are used to perform. To understand 'hello' is to understand that to say "Hello" is to greet someone. To understand 'ditto' is to know that to utter it is to signal agreement with a previous remark. To understand 'yes' is, roughly, to understand that uttering it in response to a question like 'Are you comfortable?' is to assert that you are comfortable. To understand 'please' is to understand that adding it to sentences of certain restricted grammatical forms indicates that your remark is to be taken as a polite request. Because the meanings of these words *are* given in terms of the linguistic performances they are used to make, the range of sentences in which they can meaningfully occur is highly restricted. For example, we wouldn't normally utter any of the following examples.

[21] Why is it to praise or recommend? Because it is part of the meaning of 'praise' to ascribe a desirable property to an individual is to praise that individual, and it is part of the meaning of 'recommend' that to ascribe a property that enhances the individual's ability to fulfill the expectations for the position to be filled is to recommend the individual. Examples like these abound. To say that something, or someone, is dangerous is, normally, to *warn* someone. Nevertheless, 'dangerous' is still a descriptive term.

9a. I believe that hello.
b. If hello, then one is friendly.
c. I doubt whether ditto.
d. If ditto, then there is nothing to argue about.
e. Sam disputed Mark's claim that would you please pass the pepper.
f. I wonder whether yes.
g. If yes, then there is an even prime number.

In some cases we can force a comprehensible interpretation onto these deviant sentences, as in the following dialog:

A: Is 2 a prime number? B: If yes, then there is an even prime number.

But even here, the response would more properly be expressed "If the answer is 'yes', then there is an even prime number."

The general point remains. Since these special words have nondescriptive, performative meanings that are given by specifying the linguistic acts they are used to perform, the range of linguistic environments in which they can meaningfully occur is severely restricted. If evaluative terms were similarly nondescriptive and performative, we should expect the range of linguistic environments in which they can meaningfully occur to be similarly restricted. Since evaluative words are *not* restricted in this way, they *don't* have the meanings that the emotivist theory ascribes to them.

3.4. Revisionary Conceptions of Emotivism

If this line of argument is correct, then emotivism must be rejected as a *descriptive* theory of what evaluative words really mean in ordinary language. There is, however, another way in which it might be understood. An emotivist might maintain that our ordinary use evaluative language is confused and misguided. On the one hand, we use simple sentences containing evaluative terms to give orders, make recommendations, and generally to guide action. On the other hand, we use evaluative terms in a broader class of sentences in a different, quasi-descriptive, way—as if they were simply words standing for properties that things might have or lack. Thus, it might be claimed, our ordinary use of evaluative words presupposes both that they stand for properties of acts, individuals, and other things, and that the recognition that something has one or another of these properties is invariably motivating and action-guiding.

But this, the emotivist might maintain, is incoherent—no properties are intrinsically, and by their very nature, motivating and action-guiding. An emotivist who took this view would *reject* our ordinary evaluative notions as confused, inadequate, and ultimately inapplicable to anything.[22] In

[22] Mackie (1977) takes something like this view, without being a revisionist.

their place, she might propose that we substitute evaluative notions that really do work according to the emotivist theory. Such an emotivist would be a *revisionist*, whose aim was not to describe our existing evaluative language, but to replace it with something that was arguably preferable. Of course, one might wonder whether this is either practical or preferable. One might also wonder what could possibly make such a philosopher think that the rest of the world would follow her lead.

A potentially more telling criticism focuses on the claim that no properties are, by their very nature, action-guiding in the sense presupposed by our ordinary use of ethical and other evaluative words. Suppose, for the sake of argument, that the expression 'morally required,' as used by ordinary speakers, stands for a property P predicable of some acts. What exactly is the sense in which taking an act to have P must be action-guiding, if uses of a predicate expressing P are to be action-guiding in the way we commonly take uses of 'morally required' to be?

Consider the following proposals for capturing that sense.

AG1. For any possible rational agent whatsoever, to take an act A to have P involves, perhaps among other things, taking oneself to have a conclusive reason, grounded in one's own fundamental and unrenounceable values, to perform A.

AG2. For any possible rational agent whatsoever, to take an act A to have P involves, perhaps among other things, taking oneself to have some (not necessarily conclusive) reason, grounded in one's own fundamental and unrenounceable values, to perform A.

AG3. For actual normal human agents, to take an act A to have P involves, perhaps among other things, taking oneself to have a conclusive reason, grounded in one's own fundamental and unrenounceable values, to perform A.

AG4. For actual normal human agents, to take an act A to have P involves, perhaps among other things, taking oneself to have some (not necessarily) conclusive reason, grounded in one's own fundamental and unrenounceable values, to perform A.

As intimated in section 4 of chapter 12, AG1 and AG2 seem unnecessarily strong. The sense in which our moral obligations are not conditional on renounceable interests does not require that our interests be shared by all possible rational agents. It is enough if they are grounded in human biology, in the common cross-cultural experiences of human infants in families, and in inescapable features of the human condition and our life with others. This leaves AG3 and AG4 as candidates for the criterion that a property expressed by the predicate 'morally required' has to meet in order to capture the sense in which our use of this term is action-guiding. Because we have fundamental interests other than the social ones underlying moral behavior, AG3 is still too strong. But AG4 remains in contention. Thus, the task—for the emotivist who believes that our ordinary use of

ethical language *incoherently* treats predicates like 'morally required' both as action-guiding and as expressing properties that actions can have—is, at the very least, to show either that no properties satisfy something along the lines of AG4, or that those that do can't be expressed by 'morally required'. I am not aware of such a demonstration.

4. HISTORICAL LEGACIES OF EMOTIVISM

Two general requirements have emerged from our discussion that any theory of the uses of evaluative language must satisfy.

R1. It must explain the role of reason, reflection, and logic in evaluative matters.
R2. It must explain how the use of evaluative language is related to motivation, commitment, and action.

The tension between these requirements is a central difficulty in constructing a theory of evaluative language. Descriptivist theories, which treat such language as a species of fact-stating discourse, emphasize R1, while sometimes struggling with R2. Emotivism focused on R2, while coming to grief over R1.

There is no doubt that the theories of the original emotivists were decisively refuted by the arguments brought against them. But non-cognitivism about value, broadly conceived, didn't die with emotivism. As we will see in later volumes, important versions of non-cognitivism flourished in the second half of the twentieth century. Although they too have encountered problems, the idea that there is something special about evaluative language and thought that sets them apart from ordinary fact-stating discourse, and knowledge of the world, remains a potent force in philosophy. In this respect, the historical legacy of emotivism continues.

A different historical effect, which was initially strongly felt, but which, fortunately, didn't prove to be so long-lasting, was a drastic narrowing of the focus of philosophical thought on evaluative matters. One aspect of this narrowing was the restriction of attention to a very limited range of evaluative terms—e.g., 'good', 'bad', 'right', 'wrong', 'ought', 'duty', and a few others—for which reductive analyses to a small base of emotions and preferences seemed (for a time) to be possible. For more than two decades after the advent of emotivism, philosophical discussions of value all too often gave the impression of having lost sight of the rich and nuanced character of the domain of evaluative language available for expressing judgments. To cite just a few examples of evaluative terms, we have *ought, should, good, bad, right, wrong, fair, just, unjust obligatory permissible, valuable, praiseworthy, blameworthy, justified, excusable, forgivable, rude, polite, inconsiderate, heroic, courageous, wise, prudent, decent, slovenly, slothful, beautiful, magnificent, wonderful, charming, dainty,* and *dumpy.* When one begins to appreciate the

variety of our evaluative language, one must wonder whether a single kind of analysis will work for all evaluative expressions. Whether or not, in the long run, emotivism turns out to have contained important insights, one of its worst short-term historical effects was to encourage philosophers to ignore the many differences among evaluative terms. Fortunately, there is now a much wider appreciation that moral philosophy needs—and, happily, is now receiving—a conceptual mapping of the territory covered by different classes of evaluative terms. This mapping may or may not turn out to be compatible with an essentially non-cognitivist analysis of evaluative language. But in order to provide it, one must do more than simply declare all uses of evaluative language to be emotive.

Another temporary, but historically significant, effect of emotivism was the elevation of metaethics at the expense of normative ethics. Emotivism wasn't a view about which evaluative judgments one should accept, but a doctrine about what one does when one accepts any such judgment. Hence, the dispute over emotivism was not a dispute within ethics; it was a dispute about the nature of ethics itself. Still, taking a metaethical position does not exempt one from the need to make ethical judgments, and choose among competing ethical principles. The study of these principles, and the methods for choosing among them, is known as *normative ethics*. Since even emotivists are called upon to make ethical decisions and resolve moral quandaries, one might imagine that the pursuit of normative ethics by philosophers would have continued unabated, even in the emotivist era. That didn't prove to be so.

Instead, a commitment to emotivism tended to discourage many philosophers from doing normative ethics. One of the best indications of this is found in Stevenson's influential paper, "The Emotive Meaning of Ethical Terms." After arguing that the meaning of sentences containing 'good' is primarily emotive rather than descriptive, he ends his paper with the following paragraph.

> I may add that if 'x is good' is essentially a vehicle for suggestion, it is scarcely a statement which philosophers, any more than other men, are called upon to make. To the extent that ethics predicates the ethical terms of anything, rather than explains their meaning, *it ceases to be a reflective study*. Ethical statements are social instruments. They are used in a cooperative enterprise in which we are mutually adjusting ourselves to the interests of others. Philosophers have a part, as do all men, but not a major part.[23]

Stevenson seems to be suggesting

(i) that the job of the moral philosopher is to determine the meanings of ethical terms;

[23] Stevenson (1937 [1959]), p. 281, my emphasis.

(ii) that if those meanings are emotive, then the formulation and assessment of ethical principles specifying what is right, wrong, good, or bad is not a reflective enterprise, and so is not a proper subject for philosophy.

In short, Stevenson seems to be saying that if emotivism is correct, then there is no such thing as normative ethical theory as a reflective enterprise.

Why should this be so? Perhaps it would be so if the simplest forms of emotivism were true—in which assertive utterances of ethical sentences are seen as expressions of raw emotion, with little else in the way of intelligible content. But reflective normative inquiry is not ruled out by more sophisticated versions of emotivism (or non-cognitivism generally). All of us, emotivist or not, make moral choices. In making these choices we often appeal to moral principles grounded in commitments about which we feel confident and wholehearted. But we also encounter cases in which our principles conflict with one another, or fail to give a clear result for some other reason. In these situations, we need to extrapolate from the familiar to the unfamiliar, to find a way of modifying and extending the principles we accept, which already cover many cases we feel clear about, so that they come to provide guidance for cases about which we are presently uncertain. Even if, in the end, our most basic ethical principles turn out to rest in part on personal interests and preferences about which there can be no rational argument, it is clear that reason, argument, and reflection play a large role in formulating, testing, modifying, and extending those principles. Since this sort of reasoning is the province of normative ethics, non-cognitivist metaethics shouldn't be seen as undermining normative ethics.

The way to see this most clearly is to focus on the questions "What should *I* do?" "How should *I* live?" and "What ethical principles should *I* adopt?" as opposed to the question "What ethical principles can I demonstrate anyone must adopt, no matter what his or her particular interests or preferences?" Stevenson may well have thought that the truth of any form of emotivism precluded ethical principles from being demonstrably binding on all rational agents. I have already indicated why I believe that to be an error. But even if we put that aside, there is nothing to preclude normative ethics from being a reflective enterprise that may be practiced productively by philosophers. Unfortunately, Stevenson wasn't the only non-cognitivist to suggest otherwise.

There were, however, others who continued to do illuminating work in normative ethics, even in the age of emotivism. One of these was the great anti-emotivist, anti-consequentialist, Sir David Ross, as well as his illustrious "intuitionist" predecessor, H. A. Prichard, to whom I turn in the next chapter.

CHAPTER 14

Normative Ethics and Cognitivist Metaethics in the Age of Emotivism: H. A. Prichard and W. D. Ross

1. THE OXFORD INTUITIONISTS

In this chapter I take up two important philosophers who resisted the rising tide of non-cognitivism in metaethics and the abandonment of normative ethics by leading noncognitivists in the 1930s, while also resisting the still influential consequentialist, or "ideal utilitarian," theories of their illustrious predecessors, Henry Sidgwick and G. E. Moore. The older of the two, H. A. Prichard, was born in 1871. He entered New College Oxford in 1890 as an undergraduate on a mathematics scholarship, and ended up with double firsts in Mathematics and "Greats"—ancient history and philosophy. From 1898 to 1923, he held a fellowship at Trinity College. In 1928, he was elected White's Professor of Moral Philosophy and became of fellow of Corpus Christi College. In 1937, he retired, but continued to write and work on philosophy until his death in 1947. The second Oxford intuitionist, W. D. Ross, was born in 1877. In 1895 he graduated from Edinburgh

University with a first-class degree in classics. He then went to Balliol College, Oxford, where he earned first-class degrees in Classics and "Greats" in 1898 and 1900 respectively. After holding a fellowship at Merton College, he became a tutor and Fellow at Oriel, where he remained until he joined the army in 1915. After the war, he returned to Oxford, where he served as White's Professor of Moral Philosophy from 1923 to 1928, while the holder of the Chair was ill. When the position became vacant he refused to be a candidate, in part because he thought Prichard was the better moral philosopher of the two, and in part because he wished to devote considerable energy to other areas of philosophy, including his Aristotle scholarship, for which he was widely known. He was Provost of Oriel College from 1929 to his retirement in 1947, when he also brought out an edited volume, *Moral Obligation*, of Prichard's mostly unpublished writings. Ross also served as President of the British Academy from 1936 to 1940. He died in 1971.

The chief doctrines of Prichard and Ross were (i) that although moral properties may supervene on non-moral properties, they are not reducible to them, (ii) that moral truths correspond to moral facts, (iii) that the basis of moral knowledge is immediate, non-inferential, self-evident moral cognition, (iv) that no knowledge of the right and the good can be obtained by deductive inference from non-moral truths, (v) that morally normative claims about what one ought to do are not derivable from claims about what is good, (vi) that there is no single answer to the question *What makes right acts right?*, and (vii) that the attempt in moral philosophy to prove an encompassing truth of the sort *For all acts A, A is right/wrong, morally required/morally forbidden iff x is so-and-so* is hopeless; no such encompassing truth can be established. Though this dearth of positive conclusions about the substance of our moral lives may sound disappointing, the insights in moral epistemology, moral motivation, and moral theory to be derived from studying their work are anything but.

2. PRICHARD: DOES MORAL PHILOSOPHY REST ON A MISTAKE?

Prichard was best known for his first article, "Does Moral Philosophy Rest on a Mistake?" published in 1912. Although his answer to the provocative question is, clearly, "Yes," it takes some effort to discern exactly what the mistake is supposed to be. The article raises two deep questions about moral epistemology and moral motivation: "How can we ever know what we ought to do?" and "How does the fact that we ought to do something provide us with a reason for doing it?" Prichard thought that most moral philosophy gave the wrong answer to these questions. But what was the *fundamental mistake* which led them to do so?

Prichard raises what he takes to be the main question on the first page of the article.

> Probably to most students of Moral Philosophy there comes a time when they feel a vague sense of dissatisfaction with the whole subject. . . . It is not so much that the positions, and still more the arguments, of particular thinkers seem unconvincing, though that is true. It is rather that the aim of the subject becomes increasingly obscure. "What," it is asked, "are we really going to learn by Moral Philosophy?" "What are books on Moral Philosophy really trying to show?" . . . If we reflect on our own mental history or on the history of the subject, we feel no doubt about the nature of the demand which originates the subject. Anyone who . . . has come to feel the force of the various obligations in life, at some time or other comes to feel the irksomeness of carrying them out, and to recognize the sacrifice of interest involved; and, if thoughtful, he inevitably puts to himself the question: "Is there really a reason why I should act in the ways in which hitherto I have thought I ought to act?"[24]

Prichard asks what moral philosophy aims at. He suggests that it tries to answer the question "Is there good reason to perform the acts one has pretheoretically believed to be morally right?" But what is at issue? Is one asking whether there is good reason to think the acts one has taken to be morally obligatory really are, or is one asking whether the fact that an act is morally obligatory gives one a good reason to perform it? To ask for an answer to the first question is to ask for a proof that what one has taken to be morally required really is so. To ask for an answer to the second question is to ask for a proof that one has a reason to perform the acts that morality requires.

For Prichard, the questions are combined. He completes the passage as follows.

> May I not have been all the time under an illusion in so thinking [that I ought to have acted in certain ways]? Should not I really be justified in simply trying to have a good time? Yet, like Glaucon, feeling that somehow he ought after all to act in these ways, he asks for *proof* that this feeling is justified. In other words, he asks, "*Why* should I do these things?," and his and other people's moral philosophizing is an attempt to supply the answer—i.e. to supply by a process of reflection a proof of the truth of what he and they have prior to reflection believed immediately without proof.[25]

Prichard assumes that one who morally ought to do A has reason to do A—either because that moral obligation is its own reason, or because the feature of A that makes it morally obligatory provides the reason. He also assumes that recognition that one morally ought to do A normally carries with it a motivating reason to do A. Given this, he identifies the mistake on which moral philosophy is alleged to rest; the mistake is thinking that

[24] Prichard (1912 [2002]), p. 7.
[25] Ibid., p. 7.

it is possible to *prove*, by philosophical argument, that one morally ought to do A.

Before telling us why this is a mistake, Prichard rebuts what he takes to be two standard answers to the question "What reason does one have for doing what one has pretheoretically believed to be morally obligatory?" The first answer is roughly "Because it is in one's own enlightened self-interest." Prichard replies (i) that it isn't *always* in one's interest, and (ii) that even when it is, seeing that it is in one's interest doesn't strengthen one's confidence that the act is morally obligatory, and hence that one ought to perform it. Point (i) is clearly correct. Point (ii) requires a little more explanation.

Prichard asks, "Why should we keep our engagements to our own loss?" He imagines being given the answer "Because, when closely examined, keeping our engagements is not to our long-term disadvantage." Here is his response.

> The answer is, of course, not an answer for it fails to convince us that we ought to keep our engagements, even if successful on its own lines, it only makes us *want* to keep them. . . . But if this answer is no answer, what other can be offered? . . . Suppose, when wondering whether we really ought to act in the ways usually called moral, we are told as a means of resolving our doubt that those acts are right which produce happiness. We at once ask, "Whose happiness?" If we are told, "Our own happiness," then, though we shall lose our hesitation to act in these ways, we shall not recover our sense that we ought to do so.[26]

The fact that keeping our engagements is ultimately to our advantage isn't what makes it morally obligatory to do so. Thus, Prichard thinks, it fails to address the worry *What grounds do we have for thinking that acts we pretheoretically take to be morally required really are acts we have reason to perform because they are morally required?*

Having dismissed the first traditional answer to his question, he turns the second, which tells us

> *either* [i] that anyone's happiness is a good thing in itself, and that *therefore* we ought to do whatever will produce it, *or* [ii] that working for happiness is itself good, and the intrinsic goodness of such an action is the reason we ought to do it. The advantage of this appeal to the goodness of something consists in the fact that it avoids reference to desire, and, instead, refers to something impersonal and objective.[27]

Prichard locates the difficulty with the first form of the answer in its need to assume that it somehow makes sense to impose an obligation on a state

[26] Ibid., p. 9.
[27] Ibid., p. 9.

of affairs that it should exist, or, as he puts it, to assume that "what is good ought to be." This, he says, tacitly presupposes "that the apprehension of something good which is not an action ought to be *involves just the feeling of imperativeness or obligation which is to be aroused by the thought of the action which will originate it*. Otherwise, the argument will not lead us to *feel the obligation* to produce it by the action." This is followed by the observation that *ought* is neither an operator on a sentence, nor an impersonal predicate; it is part of a complex predicate that relates agents to acts. He says, "*the proper language is never 'So and so ought to be', but 'I ought to do so and so' . . . [for] we can only feel the imperativeness upon us of something which is in our power; for it is actions and actions alone which, directly . . . are in our power.*"[28]

Note the importance of *feeling the motivational force* that Prichard thinks must arise from one's recognition that one morally ought to perform an act. He observes that we don't recognize that doing good for arbitrary others, simply because they are agents, is morally obligatory, and so we don't feel "the imperativeness" or sense of obligation to promote it. Here, he combines the normative intuitionist thesis that "ideal utilitarianism" gives an extensionally incorrect account of right and morally obligatory action with the metaethical thesis that judgments of moral obligation carry motivational force in a way that judgments about general happiness don't. Hence, ideal utilitarianism doesn't answer Prichard's question, "What reason do I have for doing what I have pretheoretically believed I morally ought to do?"

Nor does the view that one ought to perform acts *that work for the happiness of others*. Since, for Prichard, the goodness of an action is a function of the motive behind it, he doesn't dispute that actions performed with the motive of making others happy are intrinsically good. But this doesn't make them morally obligatory. Since the motive for performing an act is not of our choosing, *what we are obligated to perform* are, he insists, simply acts, not acts with a certain motive. Thus, he concludes, all traditional answers to his question fail.

Prichard then sketches his positive views of moral obligation, moral epistemology, and moral motivation.

> The sense of obligation to do, or of the rightness of, an action of *a particular kind* is absolutely underivative or immediate. The rightness of an action consists in its being the origination of something *of a certain kind A* in a situation *of a certain kind,* a situation consisting in a certain relation B of the agent to others or to his own nature.[29]

For Prichard, we are morally obligated to perform acts of certain morally relevant types, which are known to us immediately, and non-inferentially.

[28] Ibid., pp. 9–10, my emphasis.
[29] Ibid., p. 12, my emphasis.

Since which act types one counts as performing may depend on empirical facts about the situation in which one is placed, some performances instantiate several related act types. Because of this, one may need, in order to evaluate a proposed course of action, to investigate the circumstances to see exactly what morally relevant act types one will be performing.

> To appreciate its [an act's] rightness two preliminaries may be needed. We may have to follow out the consequences of the proposed action more fully than we have hitherto done, in order to realize that in the action we should originate [a certain kind of state of affairs] A. Thus we may not appreciate the wrongness of telling a certain story until we realize that we should thereby be hurting the feelings of one of our audience. Again, we may have to take into account the relation B involved in the situation, which we had hitherto failed to notice. For instance, we may not appreciate the obligation to give X a present, until we remember that he has done us an act of kindness.[30]

Empirical investigations into the circumstances surrounding a considered action are needed to determine the morally relevant types under which a performance will fall. For this reason, our search for these types must be guided by some conception of what we are looking for. Although Prichard doesn't attempt a complete inventory of the morally relevant kinds, he does give some examples.

> The relations involved in obligations . . . are very different. . . . *The obligation to repay a benefit* involves a relation due to a past act of the benefactor. *The obligation to pay a bill* involves a relation due to a past act of ours. . . . [*T*]*he obligation to speak the truth* implies no such definite act; it involves a relation consisting in the fact that others are trusting us to speak the truth. . . . *[T]he obligation not to hurt the feelings of another* involves . . . no relation other than that involved in our both being men. . . . [W]e should admit that there is [also] *an obligation to overcome our natural timidity or greediness*, and that this involves no relations to others. Still there is a relation involved . . . to our own disposition.[31]

The first task in arriving at moral knowledge that a particular course of action is, or is not, morally required is to identify the morally relevant types it falls under. Once this is done, an immediate, non-inferential apprehension of moral obligation, which Prichard calls "moral thinking," takes over.

> [G]iven that . . . we come to recognize that the proposed act is one by which we shall originate A in a relation B, then we appreciate the obligation immediately, or directly, the appreciating being an activity of *moral* thinking. We recognize, for instance, that this performance of a service to X, who has done us a service, just in virtue of its being the performance of a service to one who

[30] Ibid., pp. 12–13, my emphasis.
[31] Ibid., p. 13, my emphasis.

> has rendered a service to the would-be agent, ought to be done by us. The apprehension is immediate in precisely the sense in which a mathematical apprehension is immediate, e.g. the apprehension that this three-sided figure, in virtue of its being, must have three angles. Both apprehensions are immediate in the sense that in both insight into the nature of the subject directly leads us to recognize its possession of the predicate; . . . [in other words] the fact apprehended is self-evident.[32]

Prichard is a foundationalist in moral epistemology. All knowledge of moral obligation rests on self-evident knowledge of the sort described. Since the propositions known are self-evident, he takes them not to be established by any proofs that moral philosophers may attempt. The comparison with mathematics suggests that he regarded foundational ethical truths to be necessary and a priori—in fact *synthetic a priori,* since (i) he took fundamental ethical terms to be indefinable, and (ii) (like Frege) he took Euclidean geometry to be synthetic a priori.

What sort of knowledge does this moral epistemology allow? Let A be an act type I am considering which I know I can perform in situation S. Let A+ be a more specific act type that I would perform if I were to perform A in S. Suppose further that for every feature F bearing on the rightness or wrongness of my performing A in S, the act type A+ incorporates F—in the sense that every possible performance of A+ in any situation would have F. Finally, suppose it is knowable a priori that A+ incorporates F. For Prichard the truth or falsity of (1a) will then depend on the truth values of (1b) and (1c).

1a. I morally ought to perform A (in situation S).
b. To perform A in situation S is to perform A+ (and to fail to perform A in S is to fail to perform A+).
c. I ought to perform A+ in situation S (if I can).

So, if I could know that (1b) and (1c) were true, I could know that (1a) was true. Since typically, I can't know (1b) a priori, I can't know (1a) a priori.

What can we know a priori? If we can know a priori what all the morally relevant features of act types are, then we should be able to have self-evident foundational knowledge of *some* statements of the form (2).

2. An act type incorporating features X, Y, Z . . . but not A, B, C . . . is right/wrong/ morally obligatory in any situation in which it can be performed.

Even if we can't know a priori what all the morally relevant features of acts are, we may have foundational a priori knowledge of *some* statements of the form 3.

[32] Ibid., pp. 12–13.

3. Any act incorporating features X, Y, etc. without incorporating any other morally relevant features is right/wrong/morally obligatory in any situation in which it can be performed.

Next consider an act type A one knows one is capable of performing in a given situation. Presumably there will be some situations of this sort in which (i) one comes to know that a performance of A will have morally relevant features, which, considered in isolation, would support a judgment that A is, or isn't, morally obligatory, (ii) one knows of no other morally relevant features that one's performance of A would have, and (iii) one has done enough investigation to be justified in thinking that there are no further morally relevant features that one's performance of A would have. In these cases we may know statements of the form (1a), though we won't know them a priori.

All-things-considered moral judgments of this sort are common. The fact that they aren't self-evident, but arise from disciplined reasoning, is one source of the alleged mistake on which Prichard takes moral philosophy to rest. Because these judgments require reasoned defense, moral philosophers may wrongly think that the need for *proof* in moral philosophy is ubiquitous. According to Prichard, it isn't.

> The plausibility of the view that obligations are not self-evident but need proof lies in the fact that an act which is referred to as an obligation may be incompletely stated. . . . If, e.g., we refer to the act of repaying X by a present merely as giving X a present, it appears, and indeed is, necessary to give a reason. In other words, wherever a moral act is regarded in this incomplete way the question "*Why* should I do it?" is perfectly legitimate. This fact suggests, but suggests wrongly, that even if the nature of the act is completely stated, it is still necessary to give a reason, or, in other words, to supply a proof.[33]

Far from being ubiquitous, the need for *moral* proof is, for Prichard, essentially nil. The needed inquiry—into relevantly more specific types under which a considered action falls—is, he thinks, entirely empirical and non-moral. The only genuinely moral cognition is the immediate, non-inferential appreciation of moral obligation, in cases in which the empirical nature of the act has been grasped completely, or completely enough.

> The negative side of all this is, of course, that we do not come to appreciate an obligation by an *argument*—i.e. by a process of non-moral thinking. And that, in particular, we do not do so by an argument of which a premiss is the ethical but not moral activity of appreciating the goodness either of the act or of a consequence of the act; i.e. that our sense of the rightness of an act is not a conclusion from our appreciation of the goodness either of *it* or anything else.[34]

[33] Ibid., p. 13.
[34] Ibid., pp. 13–14.

His point is that we don't discern moral *obligations* by moral argument, even by argument involving *ethical* premises. Rather ethics, which includes both virtue and obligation, is broader than morality, which includes only the latter. Virtues are forms of goodness, of which there are many, including performing good acts. Obligation is a different sphere; the correctness of our judgments of obligation don't depend on premises about goodness.

On Prichard's picture, one action can be an instance of several different morally relevant types, some of which support the conclusion that one ought to perform it, while others may support the conclusion that one ought not to do so (or at least the conclusion that it is not the case that one ought to perform it). "Does Moral Philosophy Rest on a Mistake?" deals with a sub-case of this sort in a footnote in which the following objection to his view is stated: "[*I]f obligations are self-evident, the problem of how we ought to act in the presence of conflicting obligations is insoluble.*"[35] Prichard responds by saying

> [O]bligation admits of degrees, and . . . where obligations conflict, the decision of what we ought to do turns not on the question "Which of the alternative courses of action will originate the greater good?" but on the question "Which is the greater obligation?"[36]

He reinforces this conclusion in 1928 in his posthumously published "Conflicts of Duty" and in his Inaugural Lecture for the White's Professorship of Moral Philosophy, published in 1929.[37] The question at issue involves apparent conflicts of duty—in cases in which one must choose between acts A and B, which, though they can't both be performed, would each, when considered on its own, be judged a duty. The conclusion Prichard wants to avoid is that we are morally required to do A *and* we are morally required to do B, even though it is impossible for us to do both. This leads him to acknowledge a complication in his moral epistemology.

> The plain fact is that in the end we get driven to conclude . . . that a conflict of [all-things-considered] moral duties must be impossible . . . that a so-called statement of moral principle, to be really defensible, must be understood as stating, not that some kind of action is a duty, but that it is something else. If we then ask ourselves what this something else is, we seem driven to say that . . . it is best described as there being a claim on us to do the action, and to say for instance that that to which our having promised to do something gives rise is, strictly speaking, not a duty but a claim on us to carry it out. . . . Hence, provided we allow, as we seem driven to do, that what are usually thought of as and are called 'duties', are really claims on us to do certain actions, then we are driven to the following general conclusion. *"In any situation we are morally*

[35] Ibid., p. 14, my emphasis.
[36] Ibid., p. 14.
[37] The latter is Prichard (1929 [2002]), the former is Prichard (1928 [2002]).

bound to do that act of those which various different circumstances severally give rise to a claim on us to do, the claim on us to do which is the greatest."[38]

The significance of this passage is *not* what it tells us about conflicts of duty—since it doesn't tell us what our obligations are in cases in which *the claims on us* to do A, and to do B, are exactly equal, while being greater than the claims on us to perform any alternative act (unless we assume that there is a disjunctive act type *doing A or B*, which is our duty). What is important is that the need to weigh the relative strengths of right- and wrong-making features of acts and their alternatives is brought front and center, thereby raising questions about the architecture of Prichard's intuitionism.

That architecture begins with a plurality of features possession of which by an act contributes to its being judged to be a right act, or one we ought to perform, and also a plurality of features possession of which by an act contributes to its being judged wrong, and not to be performed. Which features these are is, according to Prichard, a matter of direct, non-inferential ethical cognition, which we acquire by imagining or experiencing acts with those features. In this way, we come to see that certain statements—e.g., *that we ought, all other things being equal, to keep a promise,* and *that we ought not, all other things being equal, to lie*—are self-evident truths. These, along with claims about the weights of the morally relevant features of acts, and their interactions, make up the foundation of his moral epistemology. From this foundation, plus ordinary empirical knowledge, we are supposed to derive all the knowledge we have, or can have, of moral obligation.

Presumably the required statements of the weights of different morally relevant features are not, for Prichard, arrived at by philosophical argument. Since they aren't deliverances of moral theory, they too must be the contents of immediate, non-inferential moral judgments. However, it is not easy to see how these statements could be either a priori or self-evidently knowable. It is one thing to say that we have synthetic a priori insight that keeping promises, paying debts, reciprocating favors, and rewarding loyalty are, all other things being equal, things we ought to do, while lying, harming others, and betraying trust are, all other things being equal, things we ought not do. There is at least some plausibility in taking these modest claims to be self-evident, a priori truths. But think of the many, often unexpected, ways in which the plurality of right- and wrong-making features of acts may interact, sometimes conflicting with, and sometimes reinforcing, one another. The moral problems we face often involve just this sort of interaction, and the decisions we make do seem to involve some kind of weighing or comparing of these features.

[38] Prichard (1928 [2002]), p. 79, my emphasis.

But whatever such weighing or comparing involves, the results don't seem to be self-evident. Nor is it easy to see them as a priori insights into the nature of the related concepts, on a par with Prichard's examples involving the concepts of Euclidean geometry.

It is often important that one *experience,* directly or imaginatively, the conflicting or reinforcing interactions—the severity of harm, the type of betrayal, or the depth of expected loyalty. It is only by such experience that one feels the morally relevant considerations "from the inside" in a way that brings one to a decision. Among the missing ingredients provided by the experience, are, I suspect, feelings that put one in touch with the sources of one's moral motivations. To make judgments in these cases is to recognize motivationally forceful (though not always sufficient) *reasons* for acting. Because of this, an acceptable intuitionist moral epistemology requires a credible theory of moral motivation.

According to Prichard, actions are performed either from a sense of moral obligation, or from a desire for the existence of something, or both. By an act done from a sense of moral obligation he means "an action done because it is right."[39] He distinguishes this from an action the purpose of which is to satisfy a desire.

> [S]o far as we act from a sense of obligation, we have no purpose or end. By a 'purpose' or 'end' we really mean something the existence of which we desire, a desire of the existence of which leads us to act. . . . The thesis, however, that so far as we act from a sense of obligation, we have no purpose must not be misunderstood. It must not be taken either to mean or to imply that so far as we do so act we have no *motive*. . . . At bottom . . . we mean by a motive what moves us to act; a sense of obligation does sometimes move us to act. . . . Desire and the sense of obligation are co-ordinate forms or species of motive.[40]

This imagined dichotomy between two independent sources of action is difficult to understand. Since the property of being *morally right* is, for Prichard, a primitive, indefinable property, no analysis of it will illuminate its motivational force. Nor is it illuminated by the supposedly direct and unmediated inference that an act is right from premises that specify its empirically relevant features, and the "strength" of those features. Thus, it is hard to avoid the conclusion that the motivationally efficacious *sense of obligation* is, for Prichard, simply a mysterious we-know-not-what.

It should not be forgotten, however, that for him, there is, in principle, always an empirical reason why a particular right act is right. It may be right because it is the keeping of a promise, the paying of a debt, the honoring of a trust, and so on. In any such case, an agent may perform the act because the agent recognizes it to be the keeping of a promise, the

[39] Prichard (1912 [2002]), p. 14.
[40] Ibid., pp. 14–15.

payment of a debt, the honoring of a trust, and so on. Perhaps, then, to perform a right action *from the sense that it is right* is simply to perform it from *the sense that it is so-and-so,* where its being *so-and-so* is what makes it right.

Though the idea is natural, it does not appear to be what Prichard had in mind.

> [W]e must sharply distinguish morality [acting under a sense of obligation] and virtue as two independent, though related, species of goodness . . . and we must at the same time allow that it is possible to do the same act either virtuously or morally or in both ways at once. . . . An act, to be virtuous, must be done willingly or with pleasure; as such it is just not done from a sense of obligation but from some desire which is intrinsically good, as arising from some intrinsically good emotion. Thus, in an act of generosity the motive is the desire to help another arising from sympathy with that other. . . . The goodness of such an act is different from the goodness of an act to which we apply the term moral in the strict and narrow sense, viz. an act done from a sense of obligation. Its goodness lies in the intrinsic goodness of the emotion and the consequent desire under which we act, the goodness of the motive being different from the goodness of the moral motive proper, viz. the sense of duty or obligation. Nevertheless, at any rate in certain cases, an act can be done either virtuously or morally or in both ways at once. It is possible to repay a benefit either from desire to repay it, or from the feeling that we ought to do so, or from both motives combined. A doctor may tend his patients either from a desire arising out of interest in his patients . . . or from a sense of duty, or from a desire and a sense of duty combined. Further . . . we regard that action as the best in which both motives are combined.[41]

Consider a case in which I repay a debt of gratitude (and so discharge a Prichardian obligation) by bestowing a benefit in appreciation of benefits I have received. What are my motives? Well, I appreciate the benefits, including the affection and respect they conveyed. I wish to reciprocate in kind, thereby conveying my respect and admiration. Having sympathy with my benefactor, I want to do something she will value. Having formed a high opinion of her, I also want her to retain her high opinion of me, which she would not do if she thought that I had taken her for granted. All of these Prichardian motives are self- and other-regarding desires arising from emotions growing out of the action, and the relationship, that led to my incurring the obligation to reciprocate.

As far as I can tell, I have no Prichardian sense of a duty to reciprocate that is distinct from these desires and emotions. I do want to reciprocate, and I have reason to do so. But I detect no bare but motivating sense of merely wishing to do what I ought. One reason I don't is that there is no

[41] Ibid., pp. 15–16.

genuine reciprocation in this case without the sincere expression of value-laden thoughts and feelings. I also can't make out the content of Prichard's nakedly austere sense of moral obligation. I do, of course, wish to be judged to be moral, and I believe that being moral is my best strategy for bringing this about. But I also want to be to *worthy* of that judgment, and of the respect, affection, and good opinion of those I esteem that accompanies it. Since doing what I morally ought to do is a way of being worthy, doing what I ought to do is a motivating reason for me. But it doesn't seem to be Prichard's sense of moral obligation because it arises from the intertwining of my self- and other-regarding desires.

Mightn't a rational agent lack those desires? Yes, that's possible. However, normal human agents can't, I believe, be entirely free of the desires and interests from which moral behavior springs. Many individuals have moral limitations and deficiencies—including the inability to recognize some important other-regarding interests that are closely related to their more purely self-regarding ones. Because of this they sometimes fail to recognize the genuine reasons they have to do certain things they morally ought to do—with the result that they wrongly judge themselves not to be morally obligated, and/or feel no motivation to perform the required action. Would actual or possible agents whose fundamental motivating interests were devoid of other-regarding concerns be moral agents at all? If so, would they lose the rights of moral agents along with the ability to incur moral obligations? Perhaps, but Prichard didn't have to face this difficult question because he took the bare fact that an agent ought to do something to be a reason for the agent to do it. This aspect of his theory of moral epistemology and motivation was, I believe, the most serious shortcoming of his otherwise illuminating pluralist vision of the sources of our moral obligations.

3. ROSS

W. D. Ross was a contemporary of Prichard, and also of A. J. Ayer, Rudolf Carnap, and C. L. Stevenson. Unlike Ayer, Carnap, and Stevenson, he was neither a logical positivist nor an emotivist. Like Prichard, he was a cognitivist and a moral realist. He believed that ethical judgments are true or false, and that ethical truths state genuine facts. Thus, in trying to determine which moral principles we should accept, he took himself to be trying ascertain ethical truth and track moral reality. However, because his views about the factual nature of moral judgments are largely independent of his arguments about which moral principles we should adopt, one can study his normative theses without attempting to settle the question of whether his metaethical position is correct. For analytical purposes, his contribution to the normative enterprise can be divided into three parts: (i) his critique of consequentialist theories of moral obligation, (ii) his

own alternative theory of obligation, and (iii) his method of formulating and testing ethical theories. I will discuss all three.

3.1. Ross's Challenge to Consequentialism

3.1.1. WHAT IS CONSEQUENTIALISM?

Consequentialist theories of moral obligation take the rightness of an action to be completely determined by the goodness or badness of its consequences. The simplest, most general, and purest form of consequentialism is given by (C).

> C. (i) An act x is right iff there is no alternative act y open to the agent the performance of which would produce a greater balance of good over bad consequences than that produced by performing x. An act is wrong iff it isn't right. (ii) An act is obligatory iff it produces a greater balance of good over bad consequences than any other act open to the agent.

According to theories of this sort, if the state of affairs resulting from performing an act A is the best state of affairs one is able to bring about, then one morally ought to perform A. If one performs any other act which brings about a less good state of affairs, then one does something morally wrong. In short, acts are simply means to the end of bringing about the best states of affairs possible. The nature of the act itself means nothing; its only morally relevant feature is the value of the effects produced by performing it.

Different versions of consequentialism result from different decisions about what counts as good (and bad). For Ross, three things are good in themselves—virtue, knowledge, and pleasure. But his arguments against strict consequentialist theories of moral obligation, do not, for the most part, depend on precisely which things are taken to be good or bad. Except in special cases, I will not be concerned with the different theories of goodness that might be adopted in conjunction with the strict consequentialist principle. But we do need to pause for a moment over the distinction between the act performed and the consequences of performing it. It is natural, when specifying those consequences, not to include the act itself, or the fact that it has been performed, as one of the consequences. After all, the consequences of a performance of an act are things *caused* by the performance, and no performance causes itself or the fact that the act has been performed.

Although this point is often taken for granted in discussing consequentialism, sometimes it isn't. Thus, we may contrast two conceptions of consequence and consequentialism. According to *simple consequentialism* the consequences of performing an act do not include the performance itself or the fact that it was performed. Rather, the event of performing the act occurs, and then, because it has occurred, other things—its consequences—occur

later. For example, a witness at a trial lies under oath. Among the consequences of the event that is the agent's telling a lie may be that the defendant is acquitted, and that the witness is later tried for perjury. But the fact that the witness told a lie is *not* one of the consequences of the event that was the agent's lying. The second conception of consequentialism, which I will call *extended consequentialism*, differs from the first in just this respect. On this conception, the consequences of performing an act include those things caused by performance, plus the performance itself. So, in the case of the lie, the fact that the witness lied is a consequences of the lie.

The difference between these two conceptions is potentially significant because the second allows one to attach intrinsic value to the performance of the act itself, and to include this value, along with the value of the states of affairs brought about by performance, in the consequentialist calculation. This can affect whether what the agent did is characterized as right or wrong. For example, a proponent of extended consequentialism might assign events in which one lies a substantial degree of intrinsic badness, independent of the states of affairs they bring about. As a result, the "consequences" of lying, in the extended sense, would always include a substantial amount of badness, which would have to be outweighed by other good results in order for a particular case of lying not to be counted as wrong.

Like many writers on the subject, Ross didn't always distinguish between these two conceptions of consequentialism. Still, his main target seems to have been simple consequentialism—which is natural, since that is what many consequentialists themselves standardly had in mind, at least until they encountered his and similar objections. Consequently, I will take simple consequentialism to be the default consequentialist position when discussing Ross, and I will revert to extended consequentialism only when necessary.

3.1.2. CONSEQUENTIALISM IS NOT TRUE BY DEFINITION

Ross's first point, in *The Right and the Good,* is that the consequentialist principle C isn't a definition (in Moore's sense) of *right act*, *obligatory act*, or *act one ought to perform*.

> The most deliberate claim that 'right' is definable as 'productive of so and so' is made by Prof. G. E. Moore, who claims in *Principia Ethica* that 'right' means 'productive of the greatest possible good'. Now it has often been pointed out against hedonism, and by no one more clearly than Prof. Moore, that the claim that 'good' just means 'pleasant' cannot seriously be maintained; that while it may or may not be true that the only things that are good are pleasant, the statement that the good is just the pleasant is a synthetic, not an analytic proposition; that the words 'good' and 'pleasant' stand for distinct qualities, even if the things that possess the one are precisely the things that possess the other. If this were not so, it would not be intelligible that the proposition 'the good is just the pleasant' should have been maintained on the one hand, and

> denied on the other, with so much fervor; for we do not fight for or against analytic propositions; we take them for granted. Must not the same claim be made about the statement 'being right means being an act productive of the greatest good producible in the circumstances'? Is it not plain on reflection that this is not what we mean by 'right', even if it be a true statement about what is right? It seems clear for instance that when an ordinary man says it is right to fulfill promises he is not in the least thinking of the total consequences of such an act, about which he knows and cares little or nothing. 'Ideal utilitarianism' [i.e., consequentialism] is, it would appear, plausible only when it is understood not as an analysis or definition of the notion of 'right' but as a statement that all acts that are right, and only these, possess the further characteristic of being productive of the best possible consequences, and are right because they possess this other characteristic.[42]

As noted in chapter 6 of Soames (2014), Ross was right in holding that consequentialist principles like C do not qualify as Moorean definitions. Since the fact that C isn't a definition doesn't tell us anything about whether or not C is true, or acceptable, Ross's critique of consequentialism requires further argument.

His argument uses the notion of *prima facie* duties, some of which involve consequences and some of which don't. The former include what he calls *duties of beneficence*. These, he says, "rest on the mere fact that there are other beings in the world whose condition we can make better in respect of virtue, or of intelligence, or of pleasure" (which he regarded as good in themselves).[43] Rossian *prima facie* duties of *self-improvement,* which, he says, "rest on the fact that we can improve our own condition in respect of virtue or of intelligence," are also consequence-involving.[44] They are duties to produce good for oneself, along with duties of beneficence to others. In both cases, the goods we are obligated to produce are of the same sort.[45] Although Ross insists that these consequentialist considerations are relevant to determining what one ought to do, he recognizes other factors that also must be considered, including *duties not to harm others*, *duties of justice*, and *duties of special relation*.

3.1.3. DUTIES NOT TO HARM OTHERS

Regarding these duties, he says:

> I think that we should distinguish from [duties of beneficence] the duties that may be summed up under the title of 'not injuring others'. No doubt to injure

[42] Ross (1930), pp. 8–9.
[43] Ibid., p. 21.
[44] Ibid., p. 21.
[45] Ross struggled over the apparent asymmetry between the duty of producing pleasure for others, about which he expressed no doubt, and the seeming lack of any duty to produce pleasure for oneself. See ibid., pp. 24-26, where he ends up concluding we do have such duties to ourselves.

> others is incidentally to fail to do them good; but it seems to me clear that non-maleficence is apprehended as a duty distinct from that of beneficence, and as a duty of a more stringent character.[46]

Ross doesn't elaborate much on this, but it is easy to see his point. Pure consequentialist principles like C require one to treat individuals as means to the end of benefiting humanity; and, because of this, they run afoul of our duty not to harm some individuals in order to benefit others. As Ross puts it, "We should not in general consider it justifiable to kill one person in order to keep another alive, or to steal from one in order to give alms to another."[47]

We may illustrate this point by imagining a doctor with three terminally ill patients—one needing a heart transplant, one needing kidneys, and one needing a liver. We stipulate that no voluntary donors or recently deceased individuals are available, and that the only possible sources of the needed organs are healthy people with no connection to the patients, and no wish to sacrifice their lives for them. Still, the doctor realizes that her patients will die without transplants. What should she do? She could, in principle, trick a healthy person, kill him, and transplant the victim's organs in the three dying patients. There might be practical difficulties with this plan—-e.g., the need to properly match the donor with the patients in order to prevent organ rejection, the uncertainties of the operation itself, the possibility of being discovered, and so on. Let us suppose that these difficulties have been eliminated. The doctor knows a healthy person whose organs wouldn't be rejected; she knows how to kill the person without being discovered, her method of transplanting organs has a very high statistical probability of success, and she is sure that everything could be kept secret.

In such a scenario, following the gruesome plan would result in three lives saved versus one lost, whereas not following the plan would result in three lives lost. Suppose the lives of all four individuals are comparable both in their own intrinsic goodness and in the amount of good they would do for others, were they to live. Then, one naturally supposes that following the plan, and killing the one to save the three, would produce a greater balance of good consequences over bad than any alternative open to the doctor. If so, then the consequentialist principle C dictates that the doctor is *morally obligated* to that course of action. But surely, Ross thinks, that is the wrong result; not only is the doctor not obligated to kill one to save three, she is obligated *not* to do so.

Ross takes examples like these to show that the consequentialist principle C is false. In drawing this conclusion, he is both rejecting a normative principle, and interpreting that rejection from a metaethical point of view that takes moral discourse to be fact-stating. Here, we are separating those

[46] Ross (1930), p. 21.
[47] Ibid., p. 22.

ideas and considering only the first. From this perspective, one must ask whether one agrees with Ross that the doctor is *not* morally obligated to follow her murderous plan. If, as I do, one does agree, one must *reject* part (ii) of C (understood in accord with *simple consequentialism*). If one further agrees with Ross, as I do again, that it would be impermissible, and hence wrong, for the doctor to follow the plan, then one must reject part (i) of C as well. Whether or not one expresses this by calling parts (i) and (ii) of C *false* is, for present purposes, immaterial.

A dedicated consequentialist who agrees with Ross about the doctor's plan might retreat to extended consequentialism, and expand his inventory of intrinsically bad states of affairs to include any state of affairs in which someone is murdered (as opposed to simply not being saved and so allowed to die). Provided that one assigns such states a high enough degree of badness, one might get the same results as Ross does in this case. But it is not clear that this strategy of weakening consequentialism so as to accommodate Ross-type examples would work in all cases. Consider a case in which we are faced with the choice of killing an innocent person at the behest of a terrorist in order to stop him from killing three others. If, in this case, one believes that one is not morally obligated to kill the innocent party, then one may have to reject part (ii) of C, even on the extended understanding of consequences. A similar test could be applied to part (i).

As Ross sees it, the problem illustrated by our example is that principle C fails to take account of the fact that our duty not to harm innocent individuals outweighs any general duty we have to benefit others. This doesn't mean that our duty not to harm is absolute, and can never be outweighed by anything else; but it does mean that there is more to determining whether an act is right, wrong, or obligatory than impersonally tallying its consequences. One does not look *only* at the end results of an act and compare them with the end results of other possible acts. Rather, one must take into consideration how one brings about those results.

3.1.4. DUTIES OF JUSTICE

The second category of duties Ross takes to raise challenges for consequentialism are *duties of justice*. These, he says, "rest on the fact or possibility of a distribution of pleasure or happiness (or of the means thereto) which is not in accordance with the merit of the persons concerned; in such cases there arises a duty to upset or prevent such a distribution."[48] Ross has his own take on questions of the distribution of goods, and how these questions relate to consequentialism. I will approach his position indirectly. First, I will sketch sample cases involving distribution, and indicate why, from a certain commonly held perspective, they raise problems

[48] Ibid., p. 21.

for consequentialism. After that, I will examine how Ross's views about merit bear on the matter.

One challenge for consequentialist-driven distributions stems from the idea that individuals have rights, or deserve things, independent of their status as sentient beings who are potential beneficiaries of one's actions. If individuals have such rights (to life, liberty, and the like), or deserve certain things, then actions that involve unjustly depriving one of liberty, property, or something else one deserves may properly be judged to be not only non-obligatory, but also wrong—even if such actions produce some increment in the total social good that is unmatched by any alternative act open to the agent. The problem with consequentialism, from this point of view, is that it leaves no room for morally robust notions of deserving, or being entitled to, something.

The following three examples illustrate the point.

(i) A nation institutes a military draft. It is argued on consequentialist grounds that the poor should be drafted, while the productive and well-off should be exempted because (a) the latter add more, in civilian life, to the total social product than the poor do, and (b) their lives are better than those of the poor anyway—in terms of pleasure enjoyed, knowledge attained, virtue practiced, etc. Hence loss of their lives in battle would diminish the quantity of goods, as well as the collective quality of our lives, more than would the loss of the lives of the poor. Surely, this line of reasoning is wrong. Instituting a draft restricted to the poor on these grounds is *not* morally required; it is morally *prohibited*. The *prima facie* problem for consequentialism is that it neglects the fact that each person has an equal right to life and liberty.

(ii) A man works long and hard, on his own time, using only resources to which he is already entitled, to produce something to benefit of himself and his family (e.g., he builds a house). After he is finished, someone in authority correctly judges that the product of the man's labors would be enjoyed more by another family—enough so that confiscating and giving the man's work to that family would increase the total amount of good enjoyed by sentient beings as a whole more than allowing the man to keep what he created. Still, such action is neither morally obligatory, nor, arguably, even morally permissible. The *prima facie* problem for consequentialism is that it neglects the fact that, normally, goods come into the world not as manna from heaven to be distributed impartially by benevolent authorities, but as the products of human activities that give rise to rights and entitlements.

(iii) Members of group B have false beliefs about members of group A, and on that basis, strongly dislike and disapprove of them. Nevertheless, a family from group A plans to take jobs and live in a community overwhelmingly inhabited by B's. Because of the B's intense dislike of the A's, this would lead to anger, unhappiness, and unproductive resistance on

> the part of the B's that would more than offset the good that would accrue to the family of A's if they were to move in. According to consequentialism, it would seem that the family is morally obligated not to move in. But this seems transparently wrong; the unhappiness experienced by the B's should count for nothing in this case. The apparent problem for consequentialism is that it measures only the total amount of good enjoyed, not who enjoys it or why.

Are all of these unacceptable results really consequences of consequentialism? Perhaps not. In presenting the criticisms, I assumed that the consequentialist takes facts about which things are intrinsically good (or bad) to be independent of who experiences them, and how they are produced. Although this is a common view, Ross didn't share it. In chapter 2 of *The Right and the Good,* he describes duties of justice as duties to bring about "a distribution of happiness between other people *in proportion to merit*."[49] In chapter 5, he discusses the intrinsic value of pleasure and its relationship to merit.

> But reflection on the conception of merit does not support the view that pleasure is always good in itself and pain always bad in itself. For while this conception implies the conviction that pleasure when deserved is good, and pain when undeserved is bad, it also suggests strongly that pleasure when undeserved is bad and pain when deserved good.
>
> There is also another set of facts which casts doubt on the view that pleasure is always good and pain always bad. We have a decided conviction that there are bad pleasures and (though this is less obvious) that there are good pains. We think that the pleasure taken either by the agent or by a spectator in, for instance, a lustful or cruel action is bad; and we think it a good thing that people should be pained rather than pleased by contemplating vice or misery.[50]

So perhaps in case (iii) above, involving the A's and the B's, the pain, unhappiness, and general disutility that the B's would experience were the A's to move in would not, by Ross's lights, count as bad, because the B's *shouldn't* have those feelings. Given some of his general comments, Ross might even judge the pain felt by the B's to be good.

In chapter 5, Ross expands his account of intrinsic goodness to include four things, "virtue, pleasure, the allocation of pleasure to the virtuous, and knowledge (and in a less degree right opinion)."[51] According to him, pleasure is always good, except in those cases in which disqualifying characteristics are present.

> [A] state of pleasure has the property, not necessarily of being good, but of being something that is good if the state has no other characteristic that

[49] Ibid., p. 26, my emphasis.
[50] Ibid., pp. 136–37.
[51] Ibid., p. 140.

> prevents it from being good. The two characteristics that may interfere with its being good are (a) that of being contrary to desert, and (b) that of being a state which is the realization of a bad disposition.[52]

Since his theory of goodness incorporates some consideration both of desert, and of how otherwise good states of affairs arise, Ross does *not* view his duties of justice as conflicting with the general consequentialist duty to maximize the good.

> The duty of justice is particularly complicated, and the word is used to cover things which are really very different—things such as the payment of debts, the reparation of injuries done by oneself to another, and the *bringing about of a distribution of happiness between other people in proportion to merit.* I use the word to denote only the last of these three. In the fifth chapter I shall try to show that besides the three (comparatively) simple goods, virtue, knowledge, and pleasure, there is a more complex good, not reducible to these, consisting in the proportionment of happiness to virtue. The bringing of this about is a duty which we owe to all men alike. . . . *This, therefore, with beneficence and self-improvement, comes under the general principle that we should produce as much good as possible, though the good here involved is different in kind from any other.*[53]

The idea that one cannot determine which states of affairs are good, once and for all—without making some judgments about the moral character of those enjoying the good, and how that good came to be enjoyed—is powerful, and deserves more attention than I can give it here.[54] Certainly, Ross has raised an important issue. But he doesn't supply the needed details; nor, in my opinion, does he establish that our duties of justice are simply special cases of the general consequentialist duty to maximize the good. His linking of the goodness of pleasure with virtue may be sufficient to allow the consequentialist to deal with some problems of just distribution—perhaps including the third of our illustrative scenarios, involving the A's and the B's. But it is far from clear that this link resolves the problems for consequentialism posed by the first two scenarios. The problem with drafting the poor and exempting the productively well-off is not that this would upset the proper balance between virtue and happiness; the policy would be wrong even if the poor were less virtuous than others. The same may be true in the second scenario—if our hard-working producer is himself morally quite ordinary, whereas the individuals on whom the authorities wish to bestow his labors are his moral superiors. In such a case confiscation and transfer of his house might improve the general balance of happiness and virtue. But it would neither be just, morally obligatory, nor, arguably, morally permissible. What the case illustrates is

[52] Ibid., p. 138.
[53] Ibid., p. 21, my emphasis.
[54] For an interesting discussion of goodness, desert, and their relation to equality, see Kegan (1998).

that the producer has a special claim to the fruit of his labors that is not simply a function of his overall level of moral virtue. Thus, the problem for consequentialism posed by just distributions remains.

Perhaps these remaining problems for consequentialism could be solved by making the account of the good even more dependent on antecedent judgments about the justice of the process by which good things are produced and distributed. But this is highly speculative. Best, at this point, to limit ourselves to two cautious conclusions. First, questions of justice, and fair distributions, pose *prima facie* problems for consequentialism. Although some of these problems may be solvable along roughly the lines Ross suggests, it is not clear that all such problems can be handled in this way. Second, the strategy of making one's account of the good depend on one's account of moral virtue, justice, desert, entitlement, and the like is itself a major, and troubling, change in the attractive consequentialist picture of morality as a conceptually simple—even if practically difficult—maximization problem. On the standard picture, questions of goodness are settled before one attempts to resolve issues about rightness, wrongness, and the like. This simple conception of the priority of goodness falls by the wayside if, in response to the problems posed by justice, the consequentialist makes the account of goodness depend on antecedent decisions about fairness, desert, entitlement, and virtue. Since these decisions may themselves presuppose judgments about rightness, wrongness, and moral obligation, the right and the good become conceptually entangled, and the attractive simplicity of the standard consequentialist picture is destroyed. Since this conceptual simplicity has been one of its chief attractions, the threat here is real.

3.1.5. DUTIES OF SPECIAL RELATION

Ross's final criticism of consequentialism involves what may be called *duties of special relation*. These typically involve cases in which certain actions of the agent give rise to rights in other people. The existence of these rights explains why certain further acts that maximize good consequences are, nevertheless, not morally obligatory, and may not even be morally permissible.

The first such duty is to keep one's implicit and explicit promises. Ross takes lying to involve the breaking of an implicit promise one makes when one engages in a conversation.[55] To make a promise is to make a commitment to someone. Once the commitment has been made, the person to whom we have made the promise has a special claim on us that others don't have; that person no longer has the status of being simply one member of humankind who is a possible beneficiary of our action. Thus, when the time comes to do what we promised, we don't think of

[55] Ross (1930), p. 21.

maximizing good consequences for humankind as a whole, but rather of keeping our prior commitment. There may, of course, be special circumstances in which some other obligation arises that outweighs our obligation to keep our promise; for example, the need to rush my sick friend to the hospital may preclude me from keeping my promise to meet you at the movie theater. But special circumstances aside, we don't think that our obligation to keep promises is outweighed by small increments in value that may accrue to humanity in general. If we have promised to do something for x, we don't search for someone other than x who might benefit a little more from our action than x would; we simply take ourselves to be morally required to keep our original promise. Ross suggests that in recognizing this, we are, in effect, recognizing the unacceptability of strict consequentialism.

> It might seem absurd to suggest that it could be right for any one to do an act which would produce consequences less good than those which would be produced by some other act in his power. Yet a little thought will convince us that this is not absurd. The type of case in which it is easiest to see that this is so is, perhaps, that in which one has made a promise. In such a case we all think that *prima facie* it is our duty to fulfill the promise irrespective of the precise goodness of the total consequences. And though we do not think it is necessarily our actual or absolute duty to do so, we are far from thinking that any, even the slightest, gain in the value of the total consequences will necessarily justify us in doing something else instead. Suppose, to simplify the case by abstraction, the fulfillment of a promise to A would produce 1,000 units of good for him, but that by doing some other act I could produce 1,001 units of good for B, to whom I have made no promise, the other consequences of the two acts being of equal value; should we really think it self-evident that it was our duty to do the second act and not the first? I think not. We should, I fancy, hold that only a much greater disparity of value between the total consequences would justify us in failing to discharge our *prima facie* duty to A. After all, a promise is a promise, and is not to be treated so lightly as the theory we are examining would imply. What, exactly, a promise is, is not so easy to determine, but we are surely agreed that it constitutes a serious moral limitation to our freedom of action. To produce the 1,001 units of good for B rather than fulfill our promise to A would be to take, not perhaps our duty as philanthropists too seriously, but certainly our duty as makers of promises too lightly.[56]

Ross's second duty of special relation is the duty to make reparations, when one has previously injured, or otherwise wronged, someone. As in the case of promising, this duty arises from past acts of the agent that create rights in other persons. For example, if A harms an innocent person B, and later is in a position to bestow benefits, then A owes something special

[56] Ibid., pp. 34–35.

to B, even if the total effects of benefiting B are not quite as valuable as those of benefiting an uninvolved third party. Having harmed B, A has an obligation to set things right, before looking for others to benefit.

The third type of duty of special relation mentioned by Ross encompasses duties of gratitude, which arise from acceptance of benefits from others—especially if the benefits are of great value, or resulted from sacrifices by the other person. These duties are ubiquitous, and are typically owed to parents, family members, and friends.

All of these duties provide serious challenges to consequentialist principles like C. According to consequentialism, everyone who could conceivably benefit from our actions has, in principle, an equal moral claim on us. But this, it seems, is simply not so. People to whom we have made promises have a special moral claim on us to keep our promises; people whom we have harmed have a special claim on us to make restitution; benefactors—including family and friends—have a special claim on us to repay their good works. As Ross points out, the fact that consequentialism doesn't properly recognize this is one of its most glaring defects.

> The essential defect of the 'ideal utilitarian' theory [consequentialism] is that it ignores, or at least does not do full justice to, the highly personal character of duty. If the only duty is to produce the maximum of good, the question of who is to have the good—whether it is myself, or my benefactor, or a person to whom I have made a promise to confer that good on him, or a mere fellow man to whom I stand in no such special relation—should make no difference to my having a duty to produce that good. But we are all in fact sure that it makes a vast difference.[57]

If Ross is right about this, then consequentialism must be rejected, both as a theory of moral obligation, and as a theory of the moral rightness and wrongness of actions.

3.2. The Scope of Moral Obligation

Before turning to Ross's positive alternative to consequentialism, it is worth looking at a different defect with consequentialist principles like C—a defect Ross doesn't mention, but which plagues many theories, including, I will argue, his own. The defect involves the scope of moral obligation. According to principle C, every act is either obligatory or impermissible—except when the values of the total consequences of performing either of two different acts open to the agent are (a) exactly the same, and (b) not exceeded by the value of the total consequences of performing any other act open to the agent. In these rare cases C characterizes both acts as right, and neither as obligatory; in all other cases acts are classified as

[57] Ibid., p. 22.

morally wrong, and so impermissible, or morally obligatory. This is doubtful. Surely, many acts are permissible without being obligatory.

If, in my free time, I decide to read a book rather than listen to music, go to the gym rather than watch television, compose a letter to the editor of the newspaper rather than surf the internet, or start writing a new philosophy paper rather than watch the Red Sox play the Yankees, then what I do is, typically, neither obligatory nor wrong, but simply permitted. I don't have to calculate the benefits to humanity in order to determine what I ought to do; the question of obligation doesn't arise. One course of action may be better for me than another, one may be more virtuous, one may produce more long-term benefits to others than another. I might be praised, admired, or respected for doing some of these things, while being looked down upon for doing others. But that doesn't make any of these actions either morally obligatory or morally wrong.

Rather, we should recognize a distinction between acts that are morally wrong, acts that are morally permissible but not required, and acts that are morally obligatory—with the middle category of morally permissible but non-obligatory acts including a large range of acts that can be subdivided into those that are morally bad, those that are morally good, and those that are morally neutral. The former—the bad but morally permissible—include simple rudeness and lack of courtesy, some cases of failing to aid someone who has no special claim on one, even when the cost to oneself would be minimal, and some cases in which one has a right to do x, but exercising that right would be harmful to others. The latter—the morally good but non-obligatory acts—include everything from simple favors, to over-subscriptions of particular duties (doing one's duty plus a little bit more), to acts of saintliness, heroism, and self-sacrifice.

For example, I might do you a favor by giving you my ticket to the sold-out basketball game, so you can watch your favorite team. That would be mildly good from a moral point of view. But it is not my obligation to do it. If I don't give you the ticket, but attend the game myself, I won't be committing a moral wrong. Another kind of non-obligatory, but morally good, action involves doing one's duty, plus a little extra. For example, part of the job of a professor is to see students, to answer their questions, discuss their work, advise them in their studies, and so on. Suppose a professor does this and more. She converses with students during evenings and weekends by e-mail or over the phone, she lends them books and papers, and she continues to read their work and advise them after they go on to graduate school, or take up jobs of their own. Up to a point she is simply doing her duty as a teacher. But beyond that, her actions are non-obligatory, but praiseworthy and morally good. Often it is hard, if not impossible, to say precisely where duties end and acts of supererogation begin, but there is no question that there is a distinction to be made.

Finally, there are inspiring instances of saintliness, self-sacrifice, and heroism. These include the actions of figures like Albert Schweitzer and

Mother Teresa, who devoted their lives to alleviating misery, as well as those of the heroic firefighters and security officers, like Rick Rescorla, at the World Trade Center, who, after leading hundreds to safety, rushed back to the flaming towers and died attempting to rescue still others.[58] These rare individuals deserve the highest praise and admiration; they were not simply doing their duty, just as those who never rise to these heights are not, for that reason, failing to fulfill their moral obligations. One will describe them in that way only if one thinks that, except for rare instances of exact ties in the consequentialist calculus, there are just two morally significant categories of actions—those that are obligatory and those that are impermissible. But the slightest attention to the moral judgments we actually make shows that our categories of moral evaluation for actions are much richer than this. In failing to recognize this, strict consequentialist theories that incorporate C(ii) distort our moral experience.

The need for an expanded set of categories for morally evaluating actions has ramifications not only for normative theories, but also for some metaethical theories—in particular, emotivism. According to the crude version put forward by Ayer, to say that stealing is wrong is just to vent one's disapproval of stealing, and to say that helping others is right is just to express a positive attitude toward helping others. This simplistic analysis doesn't have the resources to distinguish between saying that a particular case of helping others is morally obligatory and saying that it is morally good but not required. To analyze *both* simply as expressions of one's approval would be to obliterate the distinction between the two. How a more sophisticated version of non-cognitivism might best meet this challenge is an open question.

3.3. Ross's Pluralist Theory of Moral Obligation

Ross's theory is built on the following list of morally relevant features of actions.[59]

MORALLY RELEVANT FEATURES

1. the value of the consequences of performing the act (as compared to the value of the consequences of performing other acts open to the agent)
2. whether the act is an instance of lying
3. whether the act is an instance of keeping a promise or of breaking a promise
4. whether the act is an instance of making reparations, or honoring a debt of gratitude

[58] Stewart (2002).

[59] Ross probably would not list 5 as a separate morally relevant feature, but would incorporate it under 1, as involving a special kind of goodness. I have included it as a separate feature because, as discussed above, I don't think his case for incorporating it under the heading of producing good consequences is decisive.

5. whether or not the act is just
6. whether or not the act harms others

Some of these morally relevant features are favorable, and some unfavorable. If an act has a favorable morally relevant feature, it is an instance of a *positive morally relevant kind*. If it has an unfavorable feature, it is an instance of a *negative morally relevant kind*. These two notions are used to define *prima facie duty* and *actual duty*.

Prima Facie Duty

(i) An agent has a *prima facie* duty to do x iff x is an instance of a positive morally relevant kind.

(ii) An agent has a *prima facie* duty not to do x iff x is an instance of a negative morally relevant kind.

Actual Duty

(i) An agent has a duty to do x iff x is an instance of a positive morally relevant kind and either (a) x is not an instance of any negative morally relevant kind, or (b) the *stringency* of x's positive morally relevant kinds is greater than that of x's negative morally relevant kinds.

(ii) A has a duty not to do x iff x is an instance of a negative morally relevant kind and either (a) x is not an instance of any positive morally relevant kind, or (b) the *stringency* of x's negative morally relevant kinds is greater than that of x's positive morally relevant kinds.

Although this framework is attractive, and avoids some counterexamples to consequentialism, three main causes of concern immediately present themselves. The first involves the scope of moral obligation. It seems that virtually every act will be of either a positive or a negative morally relevant kind, since whether or not the act has any of the morally relevant features corresponding to 2–6, performances of it will nearly always have consequences of some (positive or negative) value, and so receive an evaluation from feature 1. So, even if morally relevant features 2–6 don't come into play, the first feature will, by itself, generally be sufficient to generate an actual duty, thereby characterizing moral obligation as ubiquitous. Thus, Ross's theory—wrongly I think—characterizes nearly every situation as one in which we are under a moral obligation to perform some act or other (except in the presumably rare cases in which the relative stringencies of an act's positive and negative morally relevant kinds exactly cancel each other out). If so, then his theory, like consequentialism, will make a hash of our moral experience by failing to take proper account of the large and theoretically important class of permissible but non-obligatory acts.

It is clear from the following remark that Ross didn't agree.

> *It must be added, however, that if we are ever under no special obligation such as that of fidelity to a promisee or of gratitude to a benefactor, we ought to do what will*

> *produce most good*; and that even when we are under a special obligation the tendency of acts to promote general good is one of the main factors in determining whether they are right.[60]

Although I cannot accept the emphasized portion of this passage, the remainder is correct. Surely, if the value of the consequences of an act, at least for others, is great enough, one's *prima facie* duty not to lie, for example, or not to break a promise, can be overridden, thereby rendering these violations of one's *prima facie* duties permissible. Thus, consideration of the consequences of one's acts does play an important role in determining rightness, wrongness, and obligation. Contrary to Ross, I believe it also plays an important role in determining which permissible but nonobligatory acts are morally good, and which are morally bad.

It is worthwhile, in this connection, to consider a contemporaneous objection to Ross's inclusion of a requirement to maximize the good among his *prima facie* duties. The objection comes from a letter of July 14, 1932, from Prichard to Ross. In it Prichard argues that Ross's consequentialist requirement is conceptually quite different from his *prima facie* duties. Whereas the latter apply to act types in virtue of their intrinsic features, independent of their relations to other act types, the former applies to an act type iff it bears the relation *being one the performance of which would produce more good than would the performance of x* to all the other act types x the agent can perform.

> [W]hereas e.g. to describe an act as one of keeping a promise or as one of making reparation is to describe it in respect of a character it has in itself, to describe [an act] as producing as much good as possible, is only to do this *verbally*; it is really to describe it as having a character which it possesses only in relation to all the other acts the man can do. . . . This difference seems to me to be vital. . . . And the difference seems to me one which is paralleled in your distinction between some prima facie duty a man has and his duty *sans phrase*. . . . [T]he thing referred to [by the phrase 'prima facie duty'] is some character which an action of a certain kind possesses *in itself* i.e. as an instance of a certain kind and apart from its relatedness to the actions of other kinds possible to a man. . . . And I take your view about duty to be that in any given situation *the* action which it is my duty to do is that out of all the actions which I can, there is the greatest *prima facie duty* to do . . . [it is] so in virtue of a character which the act possesses only in relation to all the others. Hence while the basis of a prima facie duty is a character which the action has in itself, the basis of a particular act's being my duty is not. Hence . . . really the character of producing more good than any other possible action, while it might possibly be maintained to be the basis of some action's being *my* duty, can't be held to be a base of an action being a prima facie duty.[61]

[60] Ross (1930), p. 39, my emphasis.
[61] Prichard (2002), p. 286.

Suppose, in light of Prichard's concerns, we were to eliminate Ross's *prima facie* duty to maximize goodness. Since there would then be many situations in which agents choose among a wide variety of morally permissible actions, the scope of one's moral obligations would shrink. Nevertheless, the scope of moral assessment of one's actions wouldn't shrink, if the value of the consequences of actions were taken to contribute to their moral goodness, moral badness, or moral neutrality. However, this improvement on Ross wouldn't suffice. As he points out, we would also have to allow the value of the consequences of actions to sometimes *defeat* what would otherwise be all-things-considered duties determined by one's *prima facie* obligations. Presumably, in these cases a new duty that was at least partially determined by the value of its consequences would replace what would otherwise have been one's duty.

A theory with these features would deny Rosses implausible claim "that if we are ever under no special obligation such as that of fidelity to a promisee or of gratitude to a benefactor, we ought to do what will produce most good" while affirming his sensible observation that "when we are under a special obligation the tendency of acts to promote general good is one of the main factors in determining whether they are right."[62] However, a theory of this sort would still have to answer a difficult question. If *securing good consequences of value n, or avoiding bad consequences of value m,* are ever sufficient to substitute a new duty for what would otherwise be a non-consequentialist duty (based on the combined weight of the different *prima facie* obligations in play), how can this consequentialist feature of an act *not* be sufficient to *generate* a consequentialist duty *when no competing prima facie obligations are in play*? Perhaps the question can be answered. At any rate, it must be answered if the suggested strategy for improving Ross's moral theory is to avoid running the risk of reintroducing the same ubiquitous consequentialist obligation that led Ross himself to misrepresent our moral experience so seriously.

The second possible worry about Ross's normative theory involves how we determine which features of acts are morally relevant. Ross says it is self-evident which features are morally relevant and which are not; it is self-evident not only that producing the most good possible is *prima facie* right, but also that keeping promises, making reparations, and repaying debts of gratitude are too, while and lying and harming others are *prima facie* wrong. Although many philosophers find this appeal to self-evidence to be mysterious, it is hard to know what the alternative is. All normative theories posit some principles that don't derive support from anything more basic. Consequentialism takes fundamental claims about goodness plus the consequentialist principle C to be basic and unexplained. If one is both a cognitivist in metaethics, like Ross, and a consequentialist, then presumably one will take these principles of consequentialism

[62] Ross (1930), p. 39.

to be self-evident. If one isn't a cognitivist, then one may regard Ross's principles non-cognitively as well. Either way, some normative principles are fundamental. Although Ross takes more principles to have this status than many others do, it is far from clear that this is decisive.

There is, however, a more serious cause for concern. Since there are several positive and negative morally relevant features, one of which (involving the value of the consequences produced by performing the act) applies to virtually all acts, many acts will be instances of multiple morally relevant kinds. Worse, in virtually all interesting cases in which one looks to normative ethical theories for guidance, the acts will be instances of at least one positive morally relevant kind and at least one negative morally relevant kind. Ross's theory tells us that our actual duty in these cases is determined by the relative stringencies of those kinds. But what are their relative stringencies?

Ross says almost nothing about this. Here is his most definitive statement.

> It is worthwhile to try to state more definitely the nature of the acts that are right. . . . It is obvious that any of the acts that we do has countless effects, directly or indirectly, on countless people, and the probability is that any act, however right it be, will have adverse effects (though these may be very trivial) on some innocent people. Similarly, any wrong act will probably have beneficial effects on some deserving people. Every act therefore, viewed in some aspects, will be *prima facie* right, and viewed in others, *prima facie* wrong, and right acts can be distinguished from wrong acts only as being those which, of all those possible for the agent in the circumstances, have the greatest balance of *prima facie* rightness, in those respects in which they are *prima facie* right, over their *prima facie* wrongness, in those respects in which they are *prima facie* wrong. . . . *For the estimation of the comparative stringency of these prima facie obligations no general rules can, so far as I can see, be laid down.* We can only say that a great deal of stringency belongs to the duties of 'perfect obligation'—the duties of keeping our promises, or repairing wrongs we have done, and of returning the equivalent of services we have received. [63]

Ross, who was an eminent scholar and translator of Aristotle, completes the passage by quoting Aristotle and summing up his message.

> For the rest "the decision rests with perception." This sense of our particular duty in particular circumstances, preceded and informed by the fullest reflection we can bestow on the act in all its bearings, is highly fallible, *but it is the only guide we have to our duty.*[64]

In essence, what this remarkably pessimistic remark tells us is that a workable normative theory is impossible. If a Rossian theory doesn't specify

[63] Ibid., pp. 41–42, my emphasis.
[64] Ibid., p. 42, my emphasis.

relative stringencies of different morally relevant features of an act, it won't provide useful answers about which acts are right or wrong in most cases in which we seek guidance—since these tend to be actions about which there is both something positive and something negative to say.

Thus, we are left in an uncomfortable spot. Ross's arguments against consequentialism are powerful, and his case for multiple moral principles is persuasive. But his conclusion—that there is little or nothing that can be done to systematize our moral thinking by elaborating principles that establish priorities, and resolve conflicts between competing *prima facie* evaluations—seems to be a counsel of despair. This is disappointing. Ross didn't set out to sow the seeds of further doubt about the value of normative theory in philosophy. A man of moral and intellectual clarity, with a highly developed moral sensibility, he would have been the last person to discourage an intellectually disciplined approach to moral questions. Yet he may have done so.

He wrote at a time in which important analytic philosophers regarded normative ethics with suspicion—as something either ultimately unintelligible or not within the province of philosophy. Far from sharing their suspicions, Ross was the leading critic of what was then the main source, emotivism, of philosophical skepticism about ethics. Still, his own normative theory ended with what seemed to many to be a pessimistic conclusion about what can reasonably be expected from moral philosophy. Thus, his attempt to combat the widespread suspicion of normative ethics, and other evaluative matters, harbored by many leading analytic philosophers of his day may, inadvertently, have fed it.

3.4. Ross's Moral Methodology

I close with a word about Ross's methodology in ethics, which he describes here.

> In what has preceded, a good deal of use has been made of 'what we really think' about moral questions; a certain theory has been rejected [consequentialism, or "ideal utilitarianism"] because it does not agree with what we really think. It might be said that this is in principle wrong; that we should not be content to expound what our present moral consciousness tells us but should aim at a criticism of our existing moral consciousness in the light of theory. Now I do not doubt that the moral consciousness of men has in detail undergone a good deal of modification as regards the things we think right, at the hands of moral theory. But if we are told, for instance, that we should give up our view that there is a special obligatoriness attaching to the keeping of promises because it is self-evident that the only duty is to produce as much good as possible, we have to ask ourselves whether we really, when we reflect, *are* convinced that this is self-evident, and whether we really *can* get rid of our view that promise-keeping has a bindingness independent of productiveness

> of maximum good. In my own experience I find that I cannot . . . In fact it seems, on reflection, self-evident that a promise, simply as such, is something that *prima facie* ought to be kept, and it does *not*, on reflection, seem self-evident that production of maximum good is the only thing that makes an act obligatory. And to ask us to give up at the bidding of a theory our actual apprehension of what is right and what is wrong seems like asking people to repudiate their actual experience of beauty, at the bidding of a theory which says 'only that which satisfies such and such conditions can be beautiful'. If what I have called our actual apprehension is . . . truly an apprehension, i.e. an instance of knowledge, the request is nothing less than absurd.[65]

Ross continues,

> I would maintain, in fact, that what we are apt to describe as 'what we think' about moral questions contains a considerable amount that we do not think but know, and that *this forms the standard by reference to which the truth of any moral theory has to be tested, instead of having itself to be tested by reference to any theory*. . . . We have no more direct way of access to the facts about rightness and goodness and about what things are right or good, than by thinking about them; *the moral convictions of thoughtful and well-educated people are the data of ethics just as sense-perceptions are the data of a natural science*. Just as some of the latter have to be rejected as illusory, so have some of the former; but as the latter are rejected only when they are in conflict with other more accurate sense-perceptions, the former are rejected only when they are in conflict with other convictions which stand better the test of reflection. The existing body of moral convictions of the best people is the cumulative product of the moral reflection of many generations, which has developed an extremely delicate power of appreciation of moral distinctions; and this the theorist cannot afford to treat with anything other than the greatest respect.[66]

There are two strains here that may usefully be separated (without prejudice to the question of whether or not they are correct). The first is Ross's metaethical position of moral realism. For him, the subject matter of ethics is moral reality, just as the subject matter of natural science is physical reality; just as sense perception is the foundation of genuine knowledge of physical reality, so moral reflection, and pretheoretic moral intuition, are the foundations of genuine knowledge of moral reality.

The second strain is Ross's methodological conservatism. He takes seriously, and treats with respect, our strongest and most fundamental antecedently existing moral convictions. For Ross, there is no overturning all, or even most, of these convictions, or values, at once. We come to normative theory already having evaluative commitments that can't be

[65] Ibid., pp. 39–40.
[66] Ibid., pp. 40–41, my emphasis.

dismissed, except when they conflict with other more strongly held commitments. We can make adjustments and refinements, we can remove inconsistencies, and, in principle, we can try to modify and extend limited moral principles to which we are already committed, so that they provide defensible moral classifications of a broader range of actions, including some about which we are uncertain. In these cases, we try to formulate new principles that correctly characterize the moral status of the overwhelming majority of actions about which we are already certain, while issuing verdicts on some actions about which we are presently unsure. If we succeed, then support for the new principles provided by the antecedently clear cases will translate into support for the verdicts they issue on the previously unclear cases. In this way, we can hope to gradually increase the sphere of our moral confidence, and decrease our moral doubts. But there are limits to how far any normative theory can move us from our strongest antecedently held moral convictions.

The point is similar to the lesson derived from the Moorean point that our most basic pretheoretic convictions about what we know constitute data against which philosophical theories of knowledge must be tested. Hence no theory of knowledge—no matter how attractive it may initially appear—can be accepted if it contradicts too many of these convictions. It is also similar to a lesson drawn from the logical empiricists' failed attempt to construct a radical new theory of meaning—namely that our pretheoretic convictions about the meanings of sentences constitute data against which theories of meaningfulness are tested. Hence no such theory—no matter how initially attractive—can be correct if it contradicts too many of these pretheoretic convictions. Ross's methodological conservatism about normative theories, and his arguments against consequentialism, are examples of the same general idea.

REFERENCES

Alexander, Peter (1967). "Poincaré." In *The Encyclopedia of Philosophy*, vol. 6, ed. by Paul Edwards, New York and London: Collier Macmillan, 1967, 361–63.

Ayer, A. J. (1936 [1946]). *Language, Truth, and Logic*. 2nd ed. London: Gollancz.

—— (1936/37). "Verification and Experience." *Proceedings of the Aristotelian Society* 37:137–156; reprinted in Ayer (1959), 228–43.

—— (ed.) (1959). *Logical Positivism*. New York and London: The Free Press.

—— (1971). *Moore and Russell, The Analytical Heritage*. Cambridge, MA: Harvard University Press.

Berlin, Isaiah (1938/39). "Verification." *Proceedings of the Aristotelian Society* 39: 225–48.

Black, Max (1964). *A Companion to Wittgenstein's Tractatus*. Cambridge: Cambridge University Press.

Boolos, G. (1994). "The Advantages of Honest Toil over Theft." In Alexander George, ed., *Mathematics and Mind,* Oxford: Oxford University Press, 27–44; reprinted in Boolos (1998), 261–74.

—— (1998). *Logic, Logic, Logic*. Cambridge, MA: Harvard University Press.

Boolos, G., and R. Jeffrey (1974), *Computability and Logic*. Cambridge: Cambridge University Press.

Carnap, Rudolf (1928 [1967]). *The Logical Structure of the World and Pseudoproblems in Philosophy*, trans. by Rolf A. George. Berkeley: University of California Press, 1967. Originally published as *Der logische Aufbau der Welt*. Leipzig: Felix Meiner Verlag.

—— (1930/31 [1959]). "The Old and the New Logic." Trans. by Isaac Levi. In Ayer (1959), 133–46. Originally published as "Die alte und die neue Logik," *Erkenntnis* 1.

—— (1932 [1959]). "The Elimination of Metaphysics through Logical Analysis of Language," trans. by Arthur Pap, in Ayer (1959), 60–81. Originally published as "Uberwindung der Metaphysik durch logische Analyse der Sprache," *Erkenntnis,* 2.

—— (1932/33a). "Psychologie in physikalischer Sprache. *Erkenntnis* 3. Published as "Psychology in Physical Language," trans. by George Schick, in Ayer (1959), 165–98.

—— (1932/33b). "Über Protokollsätze," *Erkenntnis* 3:215–28.

—— (1934a). *Logische Syntax der Sprache*. Vienna: Schriften sur wissenschaftlichen Weltauffassung.

—— (1934b). "On the Character of Philosophical Problems." *Philosophy of Science* 1:5–19; reprinted in Sarkar (1996b).

——— (1935a). *Philosophy and Logical Syntax*. London: Kegan Paul, Trench, Trubner & Co.
——— (1935b). "Wahrheit und Bewahrung." *Actes du Congrès International de Philosophie Scientifique* 7 (Actualités Scientifiques et Industrielles, vol. 394), Paris: Hermann et Cie.
——— (1936/37). "Testability and Meaning." *Philosophy of Science* 3:419–71 and 4:1–40.
——— (1937). *The Logical Syntax of Language,* translation of (1934a) by Amethe Smeaton with expansions by the author. London: Routledge and Kegan Paul.
——— (1942). *Introduction to Semantics*. Cambridge, MA: Harvard University Press.
——— (1946). "Remarks on Induction and Truth." *Philosophy and Phenomenological Research* 6:590–602.
——— (1949). "Truth and Confirmation." In H. Feigl and W. Sellars, eds., *Readings in Philosophical Analysis,* New York: Appleton-Century-Crofts, 119–27.
——— (1950). "Empiricism, Semantics, and Ontology." *Revue Internationale de Philosophie* 4:20–40; revised and reprinted in Carnap (1956), 205–21.
——— (1952). "Meaning Postulates," *Philosophical Studies* 3:65–73; reprinted in Carnap (1956), 222–29.
——— (1956). *Meaning and Necessity*, 2nd edition. Chicago: University of Chicago Press.
Chisholm, Roderick (1963). "Supererogation and Offense: A Conceptual Scheme for Ethics." *Ratio* 5:1–14.
Church, Alonzo (1936a). "A Note on the Entscheidungsproblem." *Journal of Symbolic Logic* 1:40–41.
——— (1936b). "An Unsolvable Problem of Elementary Number Theory." *American Journal of Mathematics* 58:345–63.
——— (1937). "Review: A. M. Turing, On Computable Numbers, with an Application to the Entscheidungsproblem." *Journal of Symbolic Logic* 2:42–43.
——— (1944). *Introduction to Mathematical Logic*. Princeton, NJ: Princeton University Press.
——— (1949). "Review *of Language, Truth, and Logic: Second Edition*." *Journal of Symbolic Logic* 14:52–53.
Comte, Auguste (1830–42). *Cours de philosophie positive*. 1st ed., 2 vols. Paris: Rouen; republished by Bachelier (6 vols). Also published as *The Positive Philosophy of Auguste Comte*, translated and condensed by Harriet Martineau, London: J. Chapman, 1853.
Conant, James (1989). "Must We Show What We Cannot Say?" In R. Fleming and M. Payne, eds., *The Senses of Stanley Cavell,* Lewisburg, PA: Bucknell University Press, 242–83.
——— (1991). "Throwing Away the Top of the Ladder." *Yale Review* 79:328–64.
——— (2001). "Two Conceptions of *Die Uberwindung der Metaphysik*: Carnap and Early Wittgenstein." In T.G. McCarthy and S.C. Stidd, *Wittgenstein in America,* Oxford: Oxford University Press, 13–61.
——— (2002). "The Method of the *Tractatus*." In E. Reck, ed., *From Frege to Wittgenstein,* Oxford: Oxford University Press, 374–462.
Davidson, Donald (1967 [2001]). "Truth and Meaning." In Davidson, *Inquiries into Truth and Interpretation*, Oxford: Clarendon Press, 2001. Originally published in *Synthèse* 17:304–23.
Diamond, Cora (1988). "Throwing Away the Ladder: How to Read the *Tractatus*." *Philosophy* 63:5–27; reprinted in Diamond (1991), 179–204.

——— (1991). *Realistic Spirit*: *Wittgenstein, Philosophy, and the Mind*. Cambridge, MA: MIT Press.

Duhem, Pierre (1914). *La théorie physique: Son objet et sa structure*, 2nd ed. Paris: Chevalier et Rivière. Published as *The Aim and Structure of Physical Theory*, trans. by Phillip Wiener, Princeton, NJ: Princeton University Press, 1954.

Ebbs, Gary (2011). "Carnap and Quine on Truth by Convention." *Mind* 120:193–237.

Einstein, Albert (1905). "Zur Elektrodynamik bewegter Körper." *Annalen der Physik* 17:275. Published as "On the Electrodynamics of Moving Bodies," in *The Collected Papers of Albert Einstein,* vol. 2, Princeton, NJ: Princeton University Press, 1989, 275–310.

——— (1916). "Die Grundlage der allgemeinen Relativitatstheorie." *Annalen der Physik* 49:769–822. In *The Collected Papers of Albert Einstein,* vol. 6, Princeton, NJ: Princeton University Press, 1997, pp. 146–200.

Enderton, H. B. (1998). "Alonzo Church: Life and Work." *Bulletin of Symbolic Logic* 4:172.

Feinberg, Joel (1961). "Supererogation and Rules." *Ethics* 71:276–88.

Fogelin, Robert (1976). *Wittgenstein.* London and New York: Routledge.

——— (1982). "Wittgenstein's Operator *N*." *Analysis* 42:124–27.

——— (1987). *Wittgenstein,* 2nd edition. London and New York: Routledge.

Franzen, Torkel (2005). *Gödel's Theorem*. Wellesley, MA: A.K. Peters.

Friedman, Michael (1983). "Moritz Schlick's Philosophical Papers." *Philosophy of Science* 50:498–514; reprinted with a postscript in Friedman (1999), 17–43.

——— (1987). "Carnap's *Aufbau* Reconsidered." *Noûs* 21:521–45; reprinted in Friedman (1999), 89–113.

——— (1992 [1999]). "Epistemology in the *Aufbau*." In Friedman (1999), 114–51, with a postscript. Originally published in *Synthèse* 93:15–57.

——— (1999). *Reconsidering Logical Positivism.* Cambridge: Cambridge University Press, 1999.

Geach, Peter (1960). "Ascriptivism." *Philosophical Review* 69:221–25.

——— (1981). "Wittgenstein's Operator *N*." *Analysis* 41:168–71.

——— (1982). "More on Wittgenstein's Operator *N*." *Analysis* 42:127–28.

Gödel, Kurt (1930). "Die Vollstandigkeit der Axiome des logischen Funktionenkalkuls." *Monatshefte fur Mathematik und Physik* 37:349–60; trans. as "The Completeness of the Axioms of the Functional Calculus of Logic," by S. Bauer-Mengelberg, in Van Heijenoort (1967), 582–91; reprinted in Gödel 1986, pp. 102–23.

——— (1931). "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I." *Monatshefte für Mathematik und Physik,* 38:173–98; trans. as "On Formally Undecidable Propositions of 'Principia Mathematica' and Related Systems I," by Jean Van Heijenoort, in Van Heijenoort (1967), 596–616; reprinted in Gödel 1986, pp. 144–95.

——— (1932). "Über Vollständigkeit und Widerspruchsfreiheit." *Ergebnisse eines mathematischen Kolloquiums* 3:12–13. Published as "On Completeness and Consistency," in Gödel (1986), 235–37.

——— (1986). *Collected Works. I: Publications 1929–1936*. Ed. by S. Feferman, S. Kleene, G. Moore, R. Solovay, and J. van Heijenoort. Oxford: Oxford University Press.

Grunbaum, Adolf (1967). "The Philosophical Significance of Relativity Theory." In *The Encyclopedia of Philosophy,* vol. 7, ed. by Paul Edwards, New York and London: Collier Macmillan, 1967, 133–40.

Hacker, P.M.S. (2000). "Was He Trying to Whistle It?" In *The New Wittgenstein*. Alice Crary and Rupert Read, eds. London and New York: Routledge, 2000, 353–88.

Hahn, H. (1933 [1959]). "Logic, Mathematics, and the Knowledge of Nature." Trans. by Arthur Pap. In Ayer (1959), 147–61. Originally published as "Logik, Mathematik und Naturerkennen," in *Einheitswissenschaft*, no. 2, Vienna: Gerold.

Hahn, H., R. Carnap, and O. Neurath (1929). "The Scientific Conception of the World." Pamphlet, translated and reprinted in Sarkar (1996a), 321–41.

Harper, L., R. Stalnaker, and G. Pearce (1981). *Ifs*. Dordrecht: Reidel.

Hempel, Carl G. (1935). "On the Logical Positivist's Theory of Truth." *Analysis* 2:49–59.

—— (1950 [1959]). "The Empiricist Criterion of Meaning." In Ayer (1959), 108–29. Originally published in *Revue Internationale de Philosophie* 41:41–63.

Hilbert, David (1899). *Grundlagen der Geometrie*. Leipzig: Teubner. Published as *Foundations of Geometry*, trans. by L. Unger, La Salle, IL: Open Court, 1971.

Hilbert, David, and Bernays, Paul (1939). *Grundlagen der Mathematik*, vol. 2. Berlin: Springer.

Kegan, Shelly (1998). "Equality and Desert." In *What Do We Deserve?* ed. by O. McLeod and L. Pojman, Oxford: Oxford University Press, 277–97.

King, Jeff, Scott Soames, and Jeff Speaks (2014). *New Thinking about Propositions*. Oxford: Oxford University Press.

Kraft, Victor (1950). *Der Wiener Kreis*. Vienna: Springer. Published as *The Vienna Circle,* trans. by Arthur Pap, New York: Philosophical Library, 1953.

Kripke, Saul (1959). "A Completeness Theorem in Modal Logic." *Journal of Symbolic Logic* 24:1–14.

—— (1963). "Semantical Considerations on Modal Logic." *Acta Philosophica Fennica* 16:83–94.

—— (1979). "A Puzzle about Belief." In A. Margalit, ed., *Meaning and Use*, Dordrecht: Reidel, 339–83.

—— (1980). *Naming and Necessity.* Cambridge, MA: Harvard University Press.

Lewis, C. I. (1929). *Mind and the World Order*. New York: Dover.

Lewis, David (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Linsky, Leonard (1952). *Semantics and the Philosophy of Language*. Urbana: University of Illinois.

Lob, Martin (1955). "Solution to a Problem of Leon Henkin." *Journal of Symbolic Logic* 20:115–18.

Lowenheim, Leopold (1915). "Über Möglichkeiten im Relativkalkul." *Mathematischen Annalen* 76:447–70; translated and reprinted in van Heijenoort (1967), 228–51.

Mach, Ernst (1914). *The Analysis of Sensations.* Trans. from the first German edition (1886) by C. M. Williams, revised from the fifth German edition by Sydney Waterloo, Chicago and London: Open Court.

Mackie, John (1977). *Ethics*. Singapore: Pelican Books.

McGinn, Marie (2006). *Elucidating the Tractatus*. Oxford: Clarendon Press.

Montague, Richard (1974). *Formal Philosophy: Selected Papers of Richard Montague.* New Haven, CT: Yale.

Moore, G. E. (1901/2). "Mr. McTaggart's Studies in Hegelian Cosmology." *Proceedings of the Aristotelian Society* 2:177–212.

—— (1903). *Principia Ethica*. Cambridge: Cambridge University Press.

—— (1907/8). "Professor James' 'Pragmatism.'" *Proceedings of the Aristotelian Society* 8:33–77; reprinted in Moore (1922), 97–146.
—— (1919/20). "External and Internal Relations." *Proceedings of the Aristotelian Society* 20:40–62; reprinted in Moore (1922), 276–309.
—— (1922). *Philosophical Studies*. London: Routledge and Kegan Paul; reprinted in 1968 by Littlefield and Adams.
—— (1925). "A Defense of Common Sense." In J. H. Muirhead, ed., *Contemporary British Philosophy*, 2nd Series, New York: Macmillan; reprinted in G. E. Moore (1958), 32–59.
—— (1939). "Proof of an External World." *Proceedings of the British Academy* 25; reprinted in Moore (1958).
—— (1953). *Some Main Problems of Philosophy.* London: George Allen and Unwin.
Neurath, Otto (1932/33). "Protokollsatze." *Erkenntnis* 3:204–14; published as "Protocol Sentences," trans. by George Schick, in Ayer (1959), 99–208.
Poincaré, Henri (1902). *La Science et l'Hypothèse.* Paris: Flammarion. Published as *Science and Hypothesis,* trans. by W. J. Greenstreet, London: Scott, 1905.
Prichard, H. A. (1912 [2002]). "Does Moral Philosophy Rest on a Mistake?" *Mind* 21:21–37.
—— (1928 [2002]). "Conflicts of Duty." First published in Prichard (2002).
—— (1929 [2002]). "Duty and Interest." Inaugural Lecture, White's Professorship of Moral Philosophy. Oxford: Clarendon Press.
—— (2002). *Moral Writings*. Ed. by Jim MacAdam. Oxford: Clarendon Press.
Proops, Ian (2001). "The New Wittgenstein: A Critique." *European Journal of Philosophy* 9:375–404.
Quine, W.V.O. (1936). "Truth by Convention." In O. H. Lee, ed., *Philosophical Essays for A. N. Whitehead*, New York: Longmans; reprinted in Quine, *Ways of Paradox,* New York: Random House, 1966, 70–99.
—— (1951). "Two Dogmas of Empiricism," *Philosophical Review* 60:20–43.
Ramsey, F. P. (1923). "Critical Notice of the *Tractatus Logico-Philosophicus*." *Mind* 32:465–78.
—— (1931). *The Foundations of Mathematics and Other Logical Essays*. London: Routledge.
Reiber, Stephen (1992). "Understanding Synonyms without Knowing That They Are Synonymous." *Analysis* 52:224–28.
Reichenbach, Hans (1938). *Experience and Prediction*. Chicago: University of Chicago.
Robinson, R. M. (1950). "An Essentially Undecidable Axiom System." *Proceedings of the International Congress of Mathematics*, 729–30.
Ross, W. D. (1930). *The Right and the Good*. Oxford: Clarendon Press.
—— (1939). *Foundations of Ethics*. Oxford: Clarendon Press.
Rosser, J. Barkley (1937). "Gödel Theorems for Non-Constructive Logics." *Journal of Symbolic Logic* 2:129–37.
Russell, Bertrand (1914a). "On the Scientific Method in Philosophy." Originally given as the Herbert Spencer Lecture at Oxford; reprinted in Russell (1917).
—— (1914b). *Our Knowledge of the External World*. Chicago and London: Open Court.
—— (1917). *Mysticism and Logic*. London: Allen and Unwin, 1917, 97–124.
—— (1918/1919). "The Philosophy of Logical Atomism." *Monist* 5, no. 28:495–527, continued in *Monist* 5, no. 29:32–63, 190–222, 345–80. Reprinted in *The*

Philosophy of Logical Atomism, Peru, IL: Open Court Publishing, with an introduction by David Pears, 1985.

—— (1919). *Introduction to Mathematical Philosophy.* London: George Allen and Unwin; reprinted in 1993 by Dover.

—— (1924). "Logical Atomism." In *Contemporary British Philosophy*, First Series, London: George Allen & Unwin, and New York: The Macmillan Co. Reprinted in *The Philosophy of Logical Atomism*, Peru, IL: Open Court Publishing, with an introduction by David Pears, 1985.

Russell, Bertrand, and Alfred North Whitehead (1910). *Principia Mathematica*, vol. 1. Cambridge: Cambridge University Press.

—— (1912). *Principia Mathematica,* vol. 2. Cambridge: Cambridge University Press.

—— (1913). *Principia Mathematica,* vol. 3. Cambridge: Cambridge University Press.

Russell, Gillian (2008). *Truth in Virtue of Meaning*. Oxford: Oxford University Press.

Salmon, Nathan (1986). *Frege's Puzzle*. Cambridge. MIT Press.

—— (1987). "Existence." In James Tomberlin, ed., *Philosophical Perspectives* 1:49–108.

—— (1989). "How to Become a Millian Heir." *Noûs* 23:211–20.

—— (1990). "A Millian Heir Rejects the Wages of Sinn." In C. A. Anderson and J. Owens, eds., *Propositional Attitudes: The Role of Content in Logic, Language, and Mind*, Stanford, CA: CSLI, 215–47.

—— (1998). "Nonexistence," *Noûs* 32:277–319.

Sarkar, Sahotra (1996a). *The Emergence of Logical Empiricism: From 1900 to the Vienna Circle*, vol. 1. New York: Garland Publishing.

—— (1996b). *Logical Empiricism at Its Peak*, vol. 2. New York: Garland Publishing.

Schilpp, P.A. (1963). *The Philosophy of Rudolf Carnap*. La Salle, IL: Open Court.

Schlick, Moritz (1915). "Die philosophische Bedeutung des Relativsatsprinzips." *Zeitschrift für Philosophie und philosophische Kritik* 159:129–75. Published as "The Philosophical Significance of the Principle of Relativity," trans. by P. Heath, in Schlick (1978), 153–89.

—— (1917 [1978–79]). *Space and Time in Contemporary Physics*. Trans. by H. Brose. In Schlick (1978–79), vol. 1, 207–69. Originally published as *Raum und Zeit in der gegenwartigen Physik*, Berlin: Springer.

—— (1918 [1985]). *General Theory of Knowledge*. Trans. by A. Blumberg. La Salle, IL: Open Court. 1985. Originally published as *Allgemeine Erkenntnislehre,* Berlin: Naturwissenschaftliche Monographien und Lehbrucher.

—— (1930 [1939]). *Problems of Ethics*. Trans. by David Rynin. New York: Prentice Hall, 1939. Originally published as *Fragen der Ethik*, Vienna: Springer.

—— (1930/31 [1959]). "The Turning Point in Philosophy." Trans. by David Rynin. In Ayer (1959), 53–59. Originally published as "Die Wende der Philosophie," *Erkenntnis* 1.

—— (1932/33). "Positivismus und Realismus." *Erkenntnis* 3. Published as "Positivism and Realism," trans. by David Rynin, in Ayer (1959), 82–107.

—— (1934 [1959]). "The Foundation of Knowledge." Trans. by David Rynin. In Ayer (1959), pp. 209–27. Originally published as "Über das Fundament der Erkenntnis," *Erkenntnis* 4:79–99; also published as "On the Foundation of Knowledge," trans. by P. Heath, in Schlick (1978), 370–87.

—— (1978). *Moritz Schlick: Philosophical Papers,* vol. 1. Ed. by Henk L. Mulder and Barbara F.B. van de Velde-Schlick. Dordrecht: Reidel.

Schonfinkel, M. (1924). "Über die Bausteine der mathematischen Logik." *Mathematische Annalen,* 92, 305–16. Published as "On the Building Blocks of Mathematical Logic," in Van Heijenoort (1967), 355–66.
Searle, John (1962). "Meaning and Speech Acts." *Philosophical Review* 71:423–32.
Skolem, Thoralf Albert (1920). "Logisch-kombinatorische Untersuchungen über die Erfüllbarkeit oder Beweisbarkeit mathematischer Sätze nebst einem Theoreme über dichte Mengen." *Videnskapsselskapet Skrifter, I. Matematisk-naturvidenskabelig Klasse 6,* 1–36. Translated and reprinted in van Heijenoort (1967), 254–63.
Soames, Scott (1983). "Generality, Truth Functions, and Expressive Capacity in the *Tractatus*." *Philosophical Review* 92:573–89.
—— (1986). "Substitutivity." In J. J. Thomson, ed., *Essays in Honor of Richard Cartwright,* Cambridge, MA: MIT Press, 99–132.
—— (1987). "Direct Reference, Propositional Attitudes, and Semantic Content," *Philosophical Topics* 14:47–87; reprinted in Soames (2009a).
—— (1999). *Understanding Truth*. New York: Oxford University Press.
—— (2002). *Beyond Rigidity*. New York: Oxford University Press.
—— (2003a). *Philosophical Analysis in the Twentieth Century*, vol. 1. Princeton, NJ, and Oxford: Princeton University Press.
—— (2003b). *Philosophical Analysis in the Twentieth Century*, vol. 2. Princeton, NJ, and Oxford: Princeton University Press.
—— (2003c). "Understanding Deflationism." In John Hawthorne and Dean Zimmerman, eds., *Philosophical Perspectives* 17:369–83; reprinted in Soames (2009a).
—— (2005a). "Naming and Asserting." In Zoltan Szabo, ed., *Semantics vs. Pragmatics*, New York: Oxford University Press, 356–82; reprinted in Soames (2009a).
—— (2005b). *Reference and Description*. Princeton, NJ, and Oxford: Princeton University Press.
—— (2007a). "Actually." In Mark Kalderon, ed., *Proceedings of the Aristotelian Society*, supplementary volume 81, 2007, 251–77; reprinted in Soames (2009a).
—— (2007b). "What Are Natural Kinds?" *Philosophical Topics* 35:329–42.
—— (2009a). *Philosophical Essays*, vol. 1. Princeton, NJ: Princeton University Press.
—— (2009b). *Philosophical Essays*, vol. 2. Princeton, NJ: Princeton University Press.
—— (2010a). *Philosophy of Language.* Princeton, NJ: Princeton University Press.
—— (2010b). "True At." *Analysis* 71:124–33.
—— (2010c). *What Is Meaning?* Princeton, NJ: Princeton University Press.
—— (2014). *The Analytic Tradition in Philosophy,* vol. 1. Princeton, NJ, and Oxford: Princeton University Press.
—— (2015a). "Reply to Critics of the *Analytic Tradition in Philosophy*, Volume 1." *Philosophical Studies* 172:1681–96.
—— (2015b). *Rethinking Language, Mind, and Meaning*. Princeton, NJ, and Oxford: Princeton University Press.
—— (2016). "Yes, Explanation Is All We Have: Reply to Stephen Schiffer and Ben Caplan." *Philosophical Studies.*
—— (forthcoming). "Is There a Science of Morality?"
Stalnaker, Robert (1968). "A Theory of Conditionals." *Studies in Logical Theory, American Philosophical Quarterly, Monograph Series,* No. 2. Oxford: Blackwell; reprinted in Harper (1981). 41–55.

——— (1975). Indicative Conditionals." *Philosophia* 5:269–86; reprinted in Harper (1981). 193–210.
Stewart, James B. (2002). "The Real Heroes Are Dead." *New Yorker*, Feb. 11.
Stevenson, Charles (1937 [1959]). "The Emotive Meaning of Ethical Terms." In Ayer (1959). Originally published in *Mind* 46:14–31.
Tarski, Alfred (1931 [1983]). "On Definable Sets of Real Numbers." Trans. by J. H. Woodger. In Tarski (1983), 110–42. Originally published as "Sur les ensembles définissables de nombres réels. I," *Fundamenta Mathematicae* 17:210–39.
——— (1935). "Der Wahrheitsbegriff in den formalisierten Sprachen." *Studia Philosophica* 1:261–405.
——— (1935 [1983]). "The Concept of Truth in Formalized Languages." Trans. of Tarski (1935) by J. H. Woodger. In Tarski (1983), 152–278.
——— (1936). "Über den Begriff der logischen Folgerung." *Actes du Congrès International de Philosophie Scientifique* 7 (Actualités Scientifiques et Industrielles, vol. 394), Paris: Hermann et Cie., 1–11.
——— (1936 [1983]). "On the Concept of Logical Consequence." Trans. of Tarski (1936) by J. H. Woodger. In Tarski (1983), 409–20.
——— (1944 [1952]). "The Semantic Conception of Truth and the Foundations of Semantics." Reprinted in Leonard Linsky (1952). Originally published in *Philosophy and Phenomenological Research* 4:341–76.
——— (1969). "Truth and Proof." *Scientific American,* June: 63–67.
——— (1983). *Logic, Semantics, Metamathematics*, 2nd ed., ed. by J. Corcoran. Indianapolis, IN: Hackett.
Tarski, Alfred, A. Mostowski, and R.M. Robinson (1953). *Undecidable Theories*. Amsterdam: North Holland.
Turing, Alan (1936/37). "On Computable Numbers with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society*, series 2, 42:230–65.
Van Fraassen, Bas (1980). *The Scientific Image.* Oxford: Clarendon Press.
Van Heijenoort, Jean (ed.) (1967). *From Frege to Gödel.* Cambridge, MA: Harvard University Press.
Vaught, R. L. (1974). "Model Theory before 1945." In L. Henkin, et al., eds., *Proceedings of the Tarski Symposium*, Proceedings of Symposia in Pure Mathematics, vol. 25, Providence, RI: American Mathematical Society, pp. 153–72.
——— (1986). "Tarski's Work in Model Theory." *Journal of Symbolic Logic* 51:869–82.
Waismann, Friedrich (1967). *Ludwig Wittgenstein and the Vienna Circle*. Trans. by B. McGuiness. Oxford: Basil Blackwell, 1979. Originally published as *Wittgenstein und der Wiener Kreis,* ed. by B. McGuiness, Frankfurt: Suhrkamp, 1967.
Williamson, Timothy (2000). *Knowledge and Its Limits*. Oxford: Oxford University Press.
——— (2008). *The Philosophy of Philosophy*. Malden, MA, and Oxford: Wiley-Blackwell.
Wilson, John Q. (1993). *The Moral Sense.* New York: Free Press.
Wittgenstein, Ludwig (1913). "Notes on Logic," ed. and trans. by H. T. Costello, *Journal of Philosophy* 54 (1957):230–44; reprinted in *Notebooks*, edition 1; reprinted in Potter (2009).
——— (1914–16). *Notebooks. 1914–1916*, 2nd ed. Ed. by G. H. von Wright and G.E.M. Anscombe; trans. by G.E.M. Anscombe. Oxford: Blackwell. 1979.
——— (1922 [1961]). *Tractatus Logico-Philosophicus*. Trans. by D. Pears and B. McGuinness. London: Routledge, 1961.

—— (1922 [1999]). *Tractatus Logico-Philosophicus*. Trans. by C. K. Ogden. London: Routledge and Kegan Paul, reprinted in 1999 by Dover.
—— (1929). "Some Remarks on Logical Form." *Proceedings of the Aristotelian Society*, supplementary volume 9, 162–71.
—— (1995). *Ludwig Wittgenstein: Cambridge Letters*. Ed. by B. McGuinness and G. H. von Wright. New York: Harper.
Wittgenstein, Ludwig, and Friedrich Waismann (2003). *The Voices of Wittgenstein: The Vienna Circle*. Ed. by Gordon Baker. London and New York: Routledge.
Ziff, Paul (1960). *Semantic Analysis,* Ithaca, NY: Cornell University Press.

INDEX