# Reflections on Skolem's Paradox

Timothy Bays

Logical investigations can obviously be a useful tool in philosophy. They must, however, be informed by a sensitivity to the philosophical significance of the formalism and by a generous admixture of common sense, as well as a thorough understanding both of the basic concepts and of the technical details of the formal material used. It should not be supposed that the formalism can grind out philosophical results in a manner beyond the capacity of ordinary philosophical reasoning. There is no mathematical substitute for philosophy.

Saul Kripke: Is there a Problem about Substitutional Quantification?

# Contents

# Introduction

In 1922, Thoraf Skolem published a paper entitled "Some Remarks on Axiomatized Set Theory." The paper presents a new proof of a model-theoretic result originally due to Leopold Löwenheim, and it then discusses some of the philosophical implications of this result. In the course of this latter discussion, the paper introduces a model-theoretic puzzle that has come to be known as "Skolem's Paradox." The present dissertation focuses on this paradox.

Skolem's Paradox involves a seeming conflict between two theorems of modern logic: Cantor's theorem from set theory and the Löwenheim-Skolem theorem from model theory. Cantor's theorem says that there are uncountable sets—sets that are *too big* to be put into any one-to-one correspondence with the natural numbers. The Löwenheim-Skolem theorem says that if a collection of first-order sentences has a model, then it has a model which is only countable. Skolem's paradox arises when we note that the axioms of set theory can themselves be written as first-order sentences: how can the very axioms which prove the existence of uncountable sets be satisfied by a merely countable model?

Ever since Skolem formulated this paradox, philosophers have argued that issues of deep philosophical importance hang on the paradox's resolution. Skolem himself claimed that the paradox shows that set theory provides an inadequate foundation for mathematics. Later authors have used the paradox to argue that "every set is countable from some perspective" (Wang), that substitutional quantification is equivalent to objectual quantification (Fine), or that Quine's theory of ontological reduction is hopelessly flawed (Grandy and Chihara). Most recently, Hilary Putnam has claimed that the paradox has, in his words, "profound implications for the great metaphysical dispute about realism which has always been the central dispute in the philosophy of language."

It should be fairly clear, even at this preliminary stage, that there are two different things which these kinds of arguments need to do. First, they need to show that Skolem's Paradox really is a paradox! That is, they need to show that Skolem's Paradox exposes a genuine tension between Cantor's theorem and the Löwenheim-Skolem theorem and that eliminating this tension requires a modification in our initial views about, e.g., set theory. Once they have done this, they can go ahead and explain how their more dramatic—or, at least, more explicitly philosophical—conclusions are supposed to follow.

This dissertation focuses almost exclusively on the first half of this project—i.e., the half which tries to expose an initial tension between Cantor's theorem and the Löwenheim-Skolem theorem. I argue that, even on quite naive understandings of set theory and model theory, there is no such tension. Hence, Skolem's

Paradox is not a genuine paradox, and there is very little reason to worry about (or even to investigate) the more extreme consequences that are supposed to follow from this paradox.

The heart of my solution to Skolem's Paradox can be found in chapter 1. In 1.1 and 1.2, I formulate a relatively simple version of this paradox. In doing so, I attempt to disentangle the different roles which set theory, model theory and philosophy play in the paradox. I also isolate the features of the paradox which make it *feel* paradoxical in the first place.[1] Finally, in 1.3, I explain why this formulation of Skolem's Paradox does not constitute a genuine "paradox" after all.

Very roughly, my explanation goes as follows. In 1.2.1, I isolate six sentences (or, in two cases, formulas) that live in the near neighborhood of the English phrase "$x$ is uncountable." In 1.2.2, I show that Skolem's Paradox turns on an equivocation between two of these sentences. Then, in 1.3, I present a series of four arguments which show why this equivocation cannot be justified. At the end of the chapter, therefore, I conclude that this version of Skolem's Paradox exposes no real tensions between Cantor's theorem and the Löwenheim-Skolem theorem.

One way to avoid the argument of chapter 1 would be by reformulating Skolem's Paradox so as to rely on more sophisticated mathematics—i.e., not just on the Löwenheim-Skolem theorem. In chapter 2, I consider three examples of this strategy: one which involves transitive models, one which involves elementary submodels of $V_\kappa$ for some inaccessible $\kappa$, and one which involves theorems about permutations of the set-theoretic universe.[2] In the long run, I argue that all of these formulations fail and that, although it may take some mathematical drudge-work to discover this fact, they fail for the same reason that the original formulation of Skolem's Paradox fails. At the end of the day, then, I conclude that there is nothing even in the *vicinity* of Skolem's Paradox which exposes a genuine tension between Cantor's theorem and the Löwenheim-Skolem theorem.

In the course of this argument in chapter 2, another interesting feature of Skolem's Paradox comes to light. Philosophers have often argued that Skolem's Paradox is due, in essence, to the specific way that first-order models interpret existential quantifiers. More specifically, there is *one particular* quantifier in the formal expression corresponding to the English phrase "$x$ is uncountable" which tends to get the "blame" for Skolem's Paradox. In 2.2 and 2.3, I show that this analysis is, at least partially, misguided. There are many different symbols which can "take the blame" for Skolem's Paradox, and it's technically important to understand which ones can (and cannot) do this in specific cases.

In chapter 3, I examine a second puzzle concerning Skolem's Paradox. Ever since the 20's, philosophers have known that there is a relatively simple technical solution to Skolem's Paradox. Nevertheless, philosophers have continued to find the paradox tempting and have, in many cases, claimed that it poses a genuine philosophical problem (i.e., as discussed at the beginning of this introduction). Why, then, have so many philosophers found the technical "solution" to Skolem's Paradox less than adequate?

[1] See also the first few pages of 1.3.1 for a discussion of the "feel" of Skolem's Paradox.

[2] See the relevant sections of chapter 2 for definitions of, e.g., "transitive," "elementary," and "inaccessible."

In chapter 3, I suggest two reasons for this attitude. The first involves a general suspicion of the mathematical machinery that is used to formulate the technical solution, while the second argues that the technical solution overlooks something important about the role of axiomatization in mathematics. In 3.1, I formulate a version of the first position and then give a reply to it. Basically, I claim that there *may well be* something to worry about here, but that the worries in question are so fundamental that they reduce Skolem's Paradox itself to a triviality. If these worries are well-founded, then set-theoretic problems appear immediately, and Skolem's Paradox functions as mere technical window-dressing.

In 3.2, I tackle the second position. I begin by isolating the exact role which axiomatization would *need* to play in mathematics if Skolem's Paradox were to be a genuine problem. I then give two different arguments against the idea that axioms play (or even *could play*) this particular role. In the end, I conclude that neither of the arguments against the technical solution to Skolem's Paradox can really be sustained. Since these are the only arguments which threaten my own solution to Skolem's Paradox—i.e., the solution given in chapters 1 and 2—I conclude that my solution remains persuasive.

Finally, in chapter 4, I turn away from generic formulations of Skolem's Paradox to examine Putnam's so-called "model-theoretic argument against realism." I show that Putnam's argument involves mistakes of both the mathematical and the philosophical variety, and that these two types of mistake are closely related. Along the way, I clear up some of the mutual charges of question-begging which have characterized discussions between Putnam and his critics. In the end, I conclude that Putnam's use of Skolem's Paradox involves far less originality than is sometimes supposed and that his "model-theoretic argument against realism" is simply a failure.

If there is an overarching theme to this dissertation—other than its focus on Skolem's Paradox—it would be that expressed in the quotation on the first page. Genuine philosophical results require genuine philosophical arguments. If we are willing to put enough philosophy into our formulation of Skolem's Paradox, then we are very likely to get some philosophy out. The significance (and truth) of the philosophy we get out of Skolem's Paradox, however, will depend almost exclusively on the significance (and truth) of the philosophy we put in. In particular, we won't get any substantial philosophical conclusions from mere reflection on Cantor's theorem and the Löwenheim-Skolem theorem. And, by the way, there is no tension between these two theorems.

# Chapter 1

# A Simple Paradox

The first two chapters of this dissertation examine several "canonical" formulations of Skolem's Paradox. Chapter 1 focuses on an initial, quite simple, formulation of the paradox, while chapter 2 examines three more sophisticated formulations. Throughout these two chapters, my principal goal is to disentangle the roles that set theory, model theory, and philosophy play in the formulation of Skolem's Paradox. By doing so, I hope to 1.) provide a clearer understanding of the mathematics which lies behind Skolem's Paradox and 2.) identify some of the purely philosophical assumptions upon which the paradox depends.

On the mathematical side, I show that different formulations of Skolem's Paradox rest on fundamentally different mathematical underpinnings. As a result, there is no uniform "explanation" of the paradox—i.e., there is no single technical mistake whose isolation enables us to solve Skolem's Paradox.[1] On the philosophical side, I argue that no formulation of Skolem's Paradox can be genuinely "paradoxical" unless it relies on substantial—and, in my view, almost certainly mistaken—assumptions about the relationship between model theory and ordinary mathematical English.[2]

## 1.1 A Simple Paradox I

Let me begin with a fairly simple, though not very plausible, formulation of Skolem's Paradox.

> We start with a standard, first-order axiomatization of set theory, ZFC.[3] On the assumption that this axiomatization has a model, the Löwenheim-Skolem theorems ensure that it has a

[1] That being said, many presentations of Skolem's Paradox *do* involve relatively straightforward technical errors (or, at least, tendentious and misleading formulations of basic mathematical results). Part of my goal, then, is to provide a clear enough exposition of the mathematics behind Skolem's Paradox that readers can recognize such errors on their own. In my view, this already goes a long way towards "solving" Skolem's Paradox.

[2] In section 1.2 I isolate these assumptions and explain *how* they contribute to Skolem's Paradox. In 1.3, I argue that these assumptions are false, and I provide an initial explanation as to *why* they are false. I do not, however, consider any of the arguments which might be put forward in *favor* of the assumptions until chapter 3.

[3] Throughout the next two chapters, I use ZFC as a "canonical" axiomatization of set theory. At points, I leave ZFC to discuss other axiomatizations—e.g., GB, NF, etc.—but these discussions are limited to the footnotes. In the long run, nothing of philosophical significance depends on the particular (first-order) axioms used in formulating Skolem's Paradox; so restricting our discussion to ZFC is a harmless expository convenience.

> countable model. Call this model $M$. Now, as ZFC $\vdash \exists x$ "*$x$ is uncountable*," there must be some $\hat{m} \in M$ such that $M \models$ "*$\hat{m}$ is uncountable*." But, since $M$ itself is only countable, there are only countably many $m \in M$ such that $M \models m \in \hat{m}$. Thus, cardinality seems to be relative: from one perspective, $\hat{m}$ is uncountable, while from another perspective, $\hat{m}$ is countable.

There are two things that we should immediately notice about this formulation of Skolem's Paradox. First, although the invocation of "relativity" and "perspective" in the last sentence is somewhat obscure, it is clearly a response to some problem which has been created in the previous sentences. Second, these previous sentences are nowhere near as explicit as we might like concerning *what* this problem is supposed to be! To make our formulation of Skolem's Paradox precise, therefore, we should begin by determining the exact nature of this initial problem.

The most natural explication of the initial problem goes something like this. From the fact that $M \models$ "*$\hat{m}$ is uncountable,*" we conclude that $\hat{m}$ really is uncountable. From the fact that there are only countably many $m \in M$ such that $M \models m \in \hat{m}$, we conclude that $\hat{m}$ is really countable. Since these two conclusions are contradictory, we have the "problem" we are looking for. Further, because each of our two conclusions involves the countability of $\hat{m}$, we have an explanation as to why *cardinality* is the property of sets that is regarded as "relative" and why *countability* is the property that is taken to depend on our "perspective."

Running with this explanation for the moment, we can reformulate Skolem's Paradox as follows. Let $M$ be a countable model for ZFC and let $\Omega(x)$ be a formula which codes the phrase "x is uncountable." As above, we use the fact that $M \models$ ZFC to obtain an element $\hat{m} \in M$ such that $M \models \Omega[\hat{m}]$. We then argue,

1. $M$ is a countable model of ZFC. (A)
2. $\Omega(x)$ says that "x is uncountable."
3. $M \models \Omega[\hat{m}]$.

$\therefore$ 4. $\hat{m}$ is uncountable.

5. If $M$ is countable, then each $m \in M$ is also countable.

$\therefore$ 6. $\hat{m}$ is countable.

Here, lines 1 and 3 follow directly from our choice of $M$ and $\hat{m}$, and line 6 follows easily from 1, 3 and 5. Our difficulties, therefore, lie in assessing the truth of premises 2 and 5 (and, indeed, in *understanding* these premises). We also need to evaluate the validity of the inference from 1–3 to 4. The rest of this section—and its successor—will focus on these tasks.

We begin with an interpretive question which effects our understanding of lines 4–6 (and, ultimately, of line 2). When we say that an element $m \in M$ "is countable," do we mean that

I. $\{x \mid x \in m\}$ is countable

or do we mean that

II. $\{x \mid M \models x \in m\}$ is countable?

Since this question effects the very *meaning* of 2 and 4–6, its answer is a prerequisite to assessing either the truth of premises 2 and 5 or the validity of the inference from 1–3 to 4.

There are two things to notice about this question. First, the two interpretations of "$m$ is countable" differ only in the way they interpret the notion of "membership" vis-a-vis the element $m$. Interpretation I assumes that we are interested in the *real* membership relation on $m$, while interpretation II assumes that we are interested in the relation which $M$ *thinks* is the membership relation on $m$—i.e., in whatever relation on $M \times M$ serves as the interpretation of "$\in$" under the interpretation function of $M$.

Second, the two interpretations share the *same* conception of countability. Under both interpretations, for instance, line 4 asserts the existence of a bijection between $\mathbb{N}$ and some set; under both interpretations, the existence of this bijection is an issue of ordinary (naive) set theory. The difference between the two interpretations concerns the appropriate *range* of this bijection: interpretation I makes the range $\{x \mid x \in \hat{m}\}$, while interpretation II makes it $\{x \mid M \models x \in \hat{m}\}$. To put this point another way, the two interpretations agree on *how* we measure the countability of a given set, but they disagree on *which* set we want to measure—i.e., which set constitutes the set of "members" of the element $m$.

This second point is crucial for a proper understanding of Skolem's Paradox. As we move along, we will encounter formulations of Skolem's Paradox which depend on various *reinterpretations* of the notion of countability. It is important, therefore, to keep such reinterpretation of "countability" distinct from the kind of interpretation at issue in I and II. To be sure, some formulations of Skolem's Paradox will involve both kinds of interpretation; but even in such cases, it will be important to keep the issues distinct so as to better understand the differing roles they play in the generation of Skolem's Paradox.

Given this, which of the two interpretations gives the best reading of argument (A)? On the one hand, interpretation I is certainly the the most natural reading of "$m$ is countable." Further, and as we shall see in sections 1.2.1 and 1.2.2, it is the only reading on which premise 2 can be made plausible without substantial modification. Nevertheless, I think it is fairly clear that interpretation II actually gives the *best* reading of "$m$ is countable" as this phrase occurs in argument (A).

There are three considerations that favor interpretation II. First, interpretation I has the effect of making premise 5 straightforwardly false. As Paul Benacerraf has noticed with respect to another formulation of Skolem's Paradox, there is absolutely no reason to think that the countability of $M$ entails that every *member* of $M$ is also countable (i.e., "countable" in the sense of interpretation I).[4] This immediate falsification of premise 5 counts heavily against using interpretation I to understand argument (A).

This problem of Benacerraf's is interesting, and I think it sheds some important light on Skolem's Paradox. To understand the problem better, we should begin by examining the following two examples. First, suppose that $\kappa$ is an inaccessible cardinal and that $N$ is a countable, elementary submodel of $V_\kappa$. In this case, even though $N$ is countable, and even though $N \models ZFC$, $N$ still contains the uncountable set $(\aleph_1)^V$ as a *member*.[5] Second, suppose that $N$ is *any* countable model for ZFC and that $X$ is any set which

[4] See [4], 102–3. As noted, Benacerraf is not discussing argument (A) itself. However, the argument he *is* discussing is sufficiently similar to (A) to make his discussion relevant.

[5] For our purposes, there are two things which are important about this example. First, the fact that $\kappa$ is inaccessible entails that the model $\langle V_\kappa, \in\rangle$ satisfies ZFC. Second, the fact that $N$ is an elementary submodel of $V_\kappa$ entails both that $N$ also satisfies

doesn't happen to be in the domain of $N$. Then, by simply substituting $X$ for some arbitrary member of $N$ and then modifying the "membership" relation on $N$ so as to respect this substitution, we obtain another model $N'$ which 1.) contains $X$, 2.) has exactly the same cardinality as $N$, and 3.) satisfies exactly the same sentences as $N$ (e.g., ZFC). If, therefore, $X$ happens to be an uncountable set, then $N'$ will be a countable model of ZFC which contains an uncountable set as a member.[6]

These examples show why premise 5 is not, as it stands, acceptable under interpretation I. There are, however, at least two ways of modifying argument (A) so as to render premise 5 acceptable without giving up on interpretation I. The first is due to Benacerraf himself. Benacerraf notes that reformulating Skolem's Paradox in terms of *transitive* models ensures that every member of a countable model is itself a countable set. In the case of argument (A), therefore, we can simply begin by choosing $M$ to be a countable, transitive model of ZFC and then replacing 5 with

$5^t$. If $M$ is countable and transitive, then each $m \in M$ is also countable.

As this new version of premise 5 is true, the modified formulation of argument (A) allows us to neatly avoid Benacerraf's problem.

Since transitive models will be important throughout the next few sections (though not, I think, for exactly the reasons that Benacerraf suggests), it is useful to pause for a moment to clarify the mathematics behind Benacerraf's proposal. To begin, we say that a set $X$ is transitive if every member of a member of $X$ is also a member of $X$ (i.e., $y \in z \in X \implies y \in X$). We say that a *model* is transitive if its domain is a transitive set and its "membership" relation is simply the real membership relation restricted to that domain.

Given this, it should be clear how transitive models help to solve Benacerraf's problem. If $M$ is a transitive model and $m$ is an element of $M$'s domain, then every member of $m$ is also a member of $M$.

ZFC and that $N$ and $V_\kappa$ agree on the identity of cardinals which have unique first-order definitions—e.g., cardinals like $\aleph_1$, $\aleph_2$ and $\aleph_\omega$. Hence, each of these (uncountable) cardinals must be an actual *member* of $N$. As a result, the countable model $N$ is literally bursting with uncountable elements.

[6] This second example uses a technical trick which will reappear several times in the next few sections, so it is useful to spend a few moments explaining it in more detail. The example depends on two theorems of model theory. First, if two models $N$ and $N'$ are *isomorphic*—i.e., if there exists a bijection $f : N \to N'$ such that for every $a, b \in N$, $a \in_N b \iff f(a) \in_{N'} f(b)$—then these models must also be *elementarily equivalent*—i.e., for every sentence $\phi$, $N \models \phi \iff N' \models \phi$.

Second, if $N$ is a model and if $f : N \to A$ is a bijection, then $f$ carries with it a canonical method for building a model which has $A$ as its domain and which is isomorphic to $N$. To obtain this model, we simply define a relation $\in_A$ on $A \times A$ as follows:

$$a \in_A a' \iff f^{-1}(a) \in_N f^{-1}(a').$$

Given this definition, $\langle A, \in_A \rangle$ is the desired model, and $f$ itself is the desired isomorphism.

Returning to the example from the text, we find that substituting $X$ for an arbitrary $\hat{n} \in N$ amounts to constructing a bijection $f : N \to (N \cup \{X\}) \setminus \{\hat{n}\}$ such that $f(\hat{n}) = X$ and $f \restriction (N \setminus \{\hat{n}\}) = \text{Id}$. Similarly, redefining $\in$ in the manner suggested above amounts to building the very model which this bijection canonically induces. Given this, claim 1 follows directly from the definition of $N'$, claim 2 follows from the fact that $f$ is a bijection, and claim 3 follows from the fact that $f$ is an isomorphism along with the fact that isomorphic models are elementarily equivalent.

Hence, $m$ can be no bigger—i.e., can have no greater cardinality—than $M$ itself. In particular, it is not possible for a countable, transitive model to contain a uncountable set as a member; this is exactly what premise $5^t$ tries to say.[7]

Unfortunately, transitive models are sometimes hard to come by. If we assume the existence of an inaccessible cardinal—as we did, for instance, in the first example of the fourth-to-last paragraph—then we can obtain such models easily.[8] Without such an assumption, however, transitive models may be hard to find. It is consistent with ZFC, for example, to accept the existence of non-transitive models of set theory while rejecting the existence of transitive ones.[9] Indeed, it is fairly easy to find consistent extensions of ZFC which are incompatible with transitivity: even if ZFC has transitive models, these extensions do not.[10]

This, then, brings us to a second technique for modifying argument (A) so as to render (some version of) premise 5 compatible with interpretation I. Like Benacerraf's transitive-model technique, this technique ensures that $M$ contains only countable sets. Unlike the transitive-model technique, this technique involves no assumptions over and above the assumption that ZFC has a model.

To employ this new technique, we begin with an *arbitrary* model $N$ such that $N \models$ ZFC. We then choose $A$ to be an arbitrary subset of $\mathcal{P}_{\omega_1}(N)$ such that $|A| = |N|$.[11] Next, we employ a trick from footnote 6. Since, $A$ and $N$ have the same cardinality, there is a canonical method for turning $A$ into a model of ZFC

[7] Another way to think about this involves noticing that interpretations I and II *coincide* for transitive $M$. That is, if $M$ is a transitive model and $m \in M$, then $\{x \mid M \models x \in m\} = \{x \mid x \in m\}$. Thus, the fact that interpretation II solves Benacerraf's problem for *arbitrary* models entails that interpretation I solves the problem for *transitive* models.

[8] The technique for obtaining such models involves a result called the "Mostowski Collapsing Lemma." This lemma allows us to take any well-founded model—i.e., any model which contains no infinite descending $\in$-chains—and find a transitive model which is isomorphic to it. Hence, if we start with an inaccessible cardinal $\kappa$ and then apply the Collapsing Lemma to some countable, elementary submodel of $V_\kappa$, we end up with a countable, transitive model of ZFC (see footnotes 5 and 6 for further background concerning this construction).

[9] Here, I use the fact that if $M$ is a transitive model of ZFC, and if $M \models \exists N$ "$N$ is a transitive model of ZFC," then $M$ must really contain some transitive model of ZFC (to use the jargon, the property "being a transitive model of ZFC" is *absolute* between $M$ and $V$). I also use the fact that every transitive model of ZFC satisfies the sentence $\exists N$ "$N$ is a model of ZFC" (since this sentence is essentially arithmetical, and transitive models get arithmetical sentences right).

Suppose, then, that there *is* a transitive model of ZFC. As an infinite descending sequence of transitive models violates the axiom of foundation, there must be a transitive model which contains no other transitive models as members (a so-called *minimal* transitive model). This model satisfies ZFC *plus* $\exists N$ "$N$ is a model of ZFC" *plus* $\neg\exists N$ "$N$ is a transitive model of ZFC". Hence, even if transitive models exist, it is consistent with ZFC + $\exists N$ "$N$ is a model of ZFC" to assume that they *don't*.

[10] Let me give two ways of obtaining obtaining such extensions. The most straightforward way involves adding a new constant $c$ to our language and then adding the sentences "$c$ is a natural number," "$c \neq 1$," "$c \neq 2$," etc. to the axioms of ZFC. The resulting theory is (by compactness) consistent; but since the constant $c$ names a non-standard natural number, the theory cannot have transitive (or even well-founded) models.

Alternately, we could let $T$ *any* consistent, axiomatizable extension of ZFC and then note that the theory $T' = T \cup \neg\,\mathrm{Con}(T)$ is still consistent but fails to have transitive models (since, in any model of $T'$, the "natural number" witnessing $\neg\,\mathrm{Con}(T)$ has to be non-standard).

[11] Here, $\mathcal{P}_{\omega_1}(N)$ is simply an abbreviation for $\{X \mid X \subset N \text{ and } |X| < \omega_1\}$. Since $N$ is infinite, we know that there are at least $|N|$ many countable subsets of $N$. Hence, $|\mathcal{P}_{\omega_1}(N)| \geq |N|$, and obtaining the desired $A$ is easy.

(and, indeed, into a model which is isomorphic to our original $N$). As a result, we wind up with a model which 1.) satisfies ZFC, 2.) has the same size as $N$, and 3.) contains only countable sets as members.[12] Since we can assume, without loss of generality, that our original $N$ was countable, this new model lets us solve Benacerraf's problem (albeit with a minor revision of premise 5[13]).

These, then, are two ways of responding to Benacerraf's problem without abandoning interpretation I. Both of them, however, require that we select a more specific model with which to work and that we rewrite premise 5 take into account the structural details of this model. In contrast, interpretation II allows us to avoid Benacerraf's problem *without* modifying argument (A): on interpretation II, the mere fact that $M$ is countable really does ensure that every $m \in M$ "is also countable." This easy avoidance of Benacerraf's problem is one reason for preferring interpretation II to interpretation I.

A second reason for preferring interpretation II is that this interpretation allows the *countability* of $M$'s domain to play a real role in argument (A). As we have just seen, any model of set theory—whether countable or uncountable—is isomorphic to a model all of whose members are countable. On interpretation I, therefore, we can formulate a (trivial) variant of argument (A) in which the cardinality of $M$ becomes a moot point: we simply let $M$ be an uncountable model containing only countable sets, and we replace premise 5 with the simple declarative

$5^d$. Every member of $M$ is countable.

Hence, since interpretation I lets us formulate argument (A) in terms of *uncountable* models, and since the countability of our models is supposed to be a key feature—indeed, *the* key feature—of Skolem's Paradox, interpretation I should be bypassed in favor of interpretation II.

This second point is crucial and needs to be emphasized. The proponent of Skolem's Paradox thinks that something important follows from the fact that the axioms of ZFC can be satisfied by a countable model. Reading their argument along the lines suggested by interpretation I has the effect of making the countability of this model irrelevant; reading it along the lines suggested by interpretation II makes the countability of this model essential. Given this, II must be the preferred interpretation of the phrase, "$\hat{m}$ is countable" in the context of argument (A).

A third reason for preferring interpretation II is that it allows us to avoid philosophical problems that are far more basic than the (mere) problems with countability which Skolem's Paradox highlights. In contrast, interpretation I leads smack into the middle of such problems. For one thing, interpretation I raises the strong possibility that lines 4–6 in argument (A) just don't make any sense. After all, the model $M$ may well

[12] It is worth noting that there is nothing special about the fact that our final model contains only *countable* sets. The same technique can be used to obtain a model all of whose members are *finite,* and a minor modification let us obtain a model all of whose members have cardinality $\kappa$, for $\kappa$ an arbitrary cardinal. In the first case, we let the domain of our model be a subset of $\mathcal{P}_\omega(N)$ rather than $\mathcal{P}_{\omega_1}(N)$; in the second, we let this domain be a subset of $\mathcal{P}_{\kappa^+}(\kappa) \setminus \mathcal{P}_\kappa(\kappa)$. (Note that we use $\kappa$ rather than $M$ in this construction, because $\mathcal{P}_{\kappa^+}(M) \setminus \mathcal{P}_\kappa(M)$ may be empty if $|M| < \kappa$).

[13] I.e., premise 5 might be replaced with something like

$5^c$. If the domain of $M$ is a subset of $\mathcal{P}_{\omega_1}(X)$ for some $X$, then each $m \in M$ is countable.

contain no sets, only things like dogs, cats, hamsters and rabbits.[14] If so, and if the terms "countable" and "uncountable" can only be applied to sets—e.g., if they are meaningless when applied to small housepets—then lines 4–6 will also be meaningless.

Even if the terms "countable" and "uncountable" *do* apply to ordinary objects (i.e., to objects that aren't sets), interpretation I still generates problems which are more basic than Skolem's Paradox. To understand these problems, we need to notice two things. First, interpretation I makes argument (A) turn on the assumption that understanding the *organizational structure* of a model enables us to understand the *internal composition* of that model's elements. More particularly, it requires the assumption that we can start with facts about the relation $\in_M$—i.e., facts about the way $M$ happens to interpret the symbol "$\in$"—and infer facts about the real membership relation on the set-theoretic universe—or, at least, facts about the real membership relation restricted to elements in the domain of $M$.[15]

Second, once this assumption about $\in_M$ and $\in$ has been made, we can generate all the "paradoxes" we want by simply comparing these two relations *directly*. To see this, suppose that $M$ is an arbitrary model of set theory and that Gandalf and Katie are two ordinary housecats; suppose also that $\Omega'(x)$ is the formula which codes "$x$ is a set" and that $\Omega''(x, y)$ is the formula which codes "$x$ is a member of $y$." Then, using a trick from footnote 6, we can build a model $M'$ and an isomorphism $f : M \to M'$ such that $f(\aleph_0^M) =$ Gandalf and $f(\aleph_1^M) =$ Katie. At the end of the day, therefore, we are in position to formulate the following "paradoxes":

1. $\Omega'(x)$ says that "$x$ is a set."

2. $M \models \Omega'[\text{Gandalf}]$.

$\therefore$ 3. Gandalf is a set.

But, 4. Gandalf is a cat, not a set.

1. $\Omega''(x, y)$ says that "$x$ is a member of $y$."

2. $M \models \Omega''[\text{Gandalf}, \text{Katie}]$.

$\therefore$ 3. Gandalf is a member of Katie.

But, 4. Katie is a cat, and cats don't have members.

Indeed, if we are interested in using this feline machinery to develop a "paradox" which is more similar to the one in argument (A), we could simply argue as follows:

1. $M$ is a countable model of ZFC.

2. $\Omega(x)$ says that "x is uncountable."

3. $M \models \Omega[\text{Katie}]$.

$\therefore$ 4. Katie is uncountable.

5. Katie is a cat, and cats don't have members.

$\therefore$ 6. Katie is not uncountable.

Thus, the assumption that models can be expected to "know" about the internal constitution of their own members—i.e., the assumption which argument (A) has to make if it is to work in accordance with interpretation I—leads to "paradoxes" involving issues that are far more basic than those raised by Skolem's

[14] Of course, $M$ will have to contain *something* other than animals, since $M$ is infinite and there are only finitely many animals in the world. Nevertheless, there is no assurance that the other elements of $M$ will be sets, nor that the particular element $\hat{m}$ will not be an animal. Besides, and as Tony Martin has reminded me, rabbits *do* breed very rapidly.

[15] So, for instance, $\in_M$ is the relation at issue when we evaluate "$M \models \Omega[\hat{m}]$" in line 2, but $\in$ is the relation at issue in (the inferred) line 3. Similarly, $\in_M$ is the relation which makes premise 5 plausible, but $\in$ is the relation which lets us infer line 6.

Paradox. In light of this fact, our primary focus on Skolem's Paradox leads us to eschew interpretation I in favor of interpretation II.[16]

These, then, are three considerations which favor interpretation II over interpretation I. Interpretation II solves Benacerraf's problem without altering our choice of models and without modifying premise 5, it makes the countability of $M$'s domain play a genuine role in argument (A), and it avoids assumptions which lead to paradoxes more basic than Skolem's Paradox itself. In light of these considerations, we can rephrase Skolem's Paradox in the following, somewhat more explicit, manner:

1. $M$ is a countable model of ZFC.
2. $\Omega(x)$ says that "x is uncountable."
3. $M \models \Omega[\hat{m}]$.

$\therefore$ 4. $\{x \mid x \in_M \hat{m}\}$ is uncountable.

5. If $M$ is countable and $m \in M$, then $\{x \mid x \in_M m\}$ is also countable.

$\therefore$ 6. $\{x \mid x \in_M \hat{m}\}$ is countable.

In this formulation, premises 1, 3 and 5 are clearly true, and the inference from 1, 3 and 5 to 6 is clearly valid. Our difficulties, therefore, involve the assessment of premise 2 and the evaluation of the inference between 1–3 and 4. The next section will focus on these two projects.

Before turning to these projects, however, it is worth noticing one "substantial" point which the present section has established: that we cannot expect an arbitrary model—even one which satisfies ZFC—to know anything about the membership relation on the real set-theoretic universe. Even in cases where the elements of a model's domain are sets, the model itself knows nothing about the set-theoretic relationships between these sets. At best, the model knows about its own internal "membership relation," and this relation may have nothing to do with the real membership relation on the set-theoretic universe (or even the restriction of this relation to the domain of $M$).[17]

## 1.2 A Simple Paradox II

In this section, I focus mostly on questions relating to premise 2. How should this premise be interpreted? Is the premise true (under some interpretations)? Is the premise strong enough to entail line 4? Before going into details, however, I want to make two preliminary points about the role premise 2 plays in argument (A).

First, if premise 2 is going to facilitate an inference from 1–3 to 4, then a lot of philosophical baggage will have to be built into the phrase "says that." At the very least, the phrase must be interpreted so as to

[16] Of course, the fact that a model cannot "know" about the internal constitution of its members should have been clear from our initial discussion of Benacerraf's problem (see, especially, the two examples on p. 6 and the material in footnotes 6 and 12). Here, I am emphasizing the fact that, even if models *could* know about the internal constitution of their members, this would not make Skolem's Paradox philosophically interesting. Instead, it would give rise to so many other paradoxes—and to so many *more basic* paradoxes—that Skolem's Paradox itself would be completely overshadowed.

[17] For more on this theme, see sections 1.3.3, 2.2, and 2.3. In 2.3, I formulate several versions of Skolem's Paradox for which this particular confusion about "membership" provides the only real explanation.

ensure that any first-order model of ZFC will respect the conception of "says that" in question. Put more concretely (and in terms of the model we are actually discussing), the conception of "says that" will have to be strong enough to ensure that the following claim holds:

$$\forall m \in M\,[M \models \Omega[m] \implies \{x \mid x \in_M m\} \text{ is uncountable}]. \qquad (\dagger)$$

Of course, this is a minimum condition. It would be good if we could replace the conditional above with a biconditional, and it would be even better if we could get more than truth-functional equivalence—e.g., if we could show that there is some deep *semantic* connection between $\Omega(x)$ and "$x$ is uncountable." The claim above, however, is all that we need to make the inference between 1–3 and 4 valid.

Second, there are strong reasons for thinking that no possible interpretation of "says that" can both support (†) and be correct. After all, the very fact that (†) leads to Skolem's paradox provides *prima facie* grounds for rejecting it. Since (†), in conjunction with some fairly straightforward mathematics, leads to an outright contradiction, we have good reason to apply *modus tollens* and reject any interpretation of "says that" which claims to support (†).

Nothing in the remainder of this section (or chapter) will try to overcome this *prima facie* objection. Instead, I will discuss several different ways of understanding $\Omega(x)$ in order to see why this formula might *seem* to say that $x$ is uncountable (and might *seem,* therefore, to support (†) ). Having done so, I will then provide a fairly detailed analysis of just how (†) goes wrong. For several different interpretations of $\Omega(x)$, I will isolate *where* $M \models \Omega[m]$ and "$m$ is uncountable" diverge, and I will explain *how* this divergence leads to a failure of (†) (and, in consequence, to the collapse of argument (A) ).

### 1.2.1 Six Interesting Sentences

In order to understand premise 2, it is useful to distinguish four different ways of understanding the formula "$\Omega(x)$." Before doing so, it is useful to consider two different sentences of ordinary English—sentences which live, so to speak, in the near vicinity of $\Omega(x)$. I begin, therefore, by laying out these six items—i.e., two sentences of ordinary English and four versions of $\Omega(x)$—and then saying a little bit about relationships between them.[18]

We can begin with the ordinary English sentence "$x$ is uncountable." If asked what this sentence means, a set theorist will say something about the lack of a bijection between $x$ and the natural numbers. If asked about the phrase "is a bijection," she might go on to talk about collections of ordered pairs satisfying certain nice properties. Finally, if asked about the term "ordered pair," she may say something about the ways one can identify ordered pairs with sets.

[18] I will often talk about four "versions" of $\Omega(x)$ rather than four ways of interpreting a single formula. For the purposes of this section, I don't think very much turns on the question of whether the semantic distinctions I draw induce genuinely different *formulas* or whether they just give a single formula different interpretations. Even in ordinary English, questions about the individuation of words which have multiple meanings (e.g., "bank") lead rapidly to complications. I see no reason, however, to think that these complications are germane to the questions at issue here.

Taking this explanatory process to its logical conclusion, we can continue to fill in the details of "$x$ is uncountable" until we obtain a single sentence (albeit a quite long and complicated sentence) which uses no phrases other than "equals," "is a member of," "not," "if. . . then," and "there is a set $y$, such that." This sentence—which, like "$x$ is uncountable," is simply a sentence of ordinary mathematical English—is the second item in our list of six. Because this sentence is (far) too long to write out explicitly, I will use "Cantor($x$)" to denote it.[19]

Given Cantor($x$), we get our third sentence by *abbreviating* the phrases "equals," "is a member of," "not," "if. . . then," and "there is a set $y$, such that" with the symbols $=, \in, \neg, \rightarrow,$ and $\exists y$. Having done so, we obtain a transcription of Cantor($x$) which uses no symbols other than the above-mentioned $=, \in, \neg, \rightarrow,$ and $\exists y$ (and, perhaps, some punctuation). For convenience, I call the resulting sentence the "plain English version of $\Omega(x)$" and denote it by $\Omega_E(x)$. Before moving on, there are two things to notice about $\Omega_E(x)$.

First, $\Omega_E(x)$ is not generated by *interpreting* some formula of first-order set theory. We do not, that is, *begin* with a string of uninterpreted first-order symbols and then *stipulate* that these symbols are to be understood in some particular way. Instead, we begin with a sentence of ordinary mathematical English, and then use a certain collection of symbols—which just happen to be commonly used in the formulation of first-order set theory—as abbreviations for terms and phrases which already occur in this sentence.[20]

Second, and as a direct consequence of the fact that $\Omega_E(x)$ is simply an abbreviation for Cantor($x$), $\Omega_E(x)$ has *exactly* the same semantics as Cantor($x$) itself. So, the symbol "$\in$" still means "is a member of," the phrase "$\exists x$" still means "there is a set $x$ such that," etc. Most importantly, the sentence $\Omega_E(x)$ is true exactly when the sentence Cantor($x$) is true—i.e., exactly when $x$ is an uncountable set.

Note that this second point *does not* entail that $\Omega_E(x)$ is semantically unproblematic. If, for instance, there are semantic problems which effect Cantor($x$) itself—e.g., some kind of vagueness or ambiguity—then these problems will also effect $\Omega_E(x)$. It *does*, however, entail that $\Omega_E(x)$ has no *special* problems which distinguish it from Cantor($x$). In particular, the mere fact that $\Omega_E(x)$ uses symbols like $=, \in$ and $\neg$ does not *introduce* semantic problems—e.g., vagueness, ambiguity or "susceptibility to reinterpretation"—which were not already present in Cantor($x$).

[19] To see the difficulties in writing Cantor($x$) explicitly, note that even a simple phrase like "$x$ is a singleton" turns into,

> There is a set $a$ such that it is not the case that if $a$ is a member of $x$, then there exists a set $b$ such that it is not the case that if $b$ is a member of $x$ then $b$ is equal to $a$.

when it is "written out" in the manner of Cantor($x$). If we move to something marginally more complicated—say, the phrase "$x$ is an ordered pair" or "f is a function"—then we get sentences the length of good size paragraphs. Finally, Cantor($x$) itself requires several pages to write down explicitly!

[20] This abbreviatory use of "first-order symbols" is quite common in mathematics. Since $\rightarrow$ is easier to write than "if. . . then," and since $\exists x$ is considerably shorter than "there exists an $x$ such that. . . ," mathematicians often conserve time (and blackboard space) by using the latter to abbreviate the former. In theory, this practice could lead to a certain kind of confusion—where, e.g., we forget which instances of "$\exists$" are symbols in our formal language and which are simply abbreviations of ordinary English. In practice, however, this seldom causes any real problems, and I have only known one logician—Yiannis Moschovakis—who tries to avoid using "$\exists$," "$\rightarrow$" and "$\neg$" as abbreviations for ordinary English.

Let us, then, leave $\Omega_E(x)$ by the wayside and move to consider three further, and somewhat more technical, ways of understanding $\Omega(x)$. First, we might consider $\Omega(x)$ as a completely *uninterpreted* string of first-order symbols. In this case, we would start with some syntactic rules governing the generation of formulas, and then notice that these rules let us generate the particular string of symbols which constitutes $\Omega(x)$. This might be the approach to take, for instance, if we wanted to investigate the proof-theoretic properties of $\Omega(x)$. For our purposes, however, the only thing to notice about this understanding of $\Omega(x)$ is that it avoids semantics altogether: $\Omega(x)$, on this understanding, is simply a syntactically interesting string of ultimately meaningless symbols. For reference purposes, I call this the "proof-theoretic version of $\Omega(x)$" and denote it by $\Omega_P(x)$.

Second, we could consider $\Omega(x)$ from the standpoint of a model theorist, giving a semantics to *some* of the symbols in $\Omega(x)$ while leaving the interpretation of other symbols as a subject for further study. If, for instance, we wanted to do first-order model theory, then we might insist that "$\neg$" function in the standard manner and that "$\exists$" and "$\in$" have a close semantic relationship (i.e., that any admissible interpretation of "$\in$" will be *joined* with an interpretation of "$\exists$" such that "$m_1 \in m_2$" only makes sense when both $m_1$ and $m_2$ are in the range of $\exists$). At the same time, we let the exact interpretation of $\in$ and $\exists$ be fairly indefinite and take this indefiniteness itself as a subject for mathematical investigation.

Since this way of thinking about $\Omega(x)$ may be somewhat unfamiliar, three comments are in order. First, understanding $\Omega(x)$ model-theoretically *does* involve giving a semantics—or, at least, a proto-semantics—to that sentence. For one thing, it gives a standard interpretation to *some* of the symbols in $\Omega(x)$—i.e., those it employs as "logical constants"—and it may place restrictions on the interpretation of others.[21] Thus, even though a model-theoretic understanding of $\Omega(x)$ does not fix a *unique* semantics for that formula, it does not leave the formula completely uninterpreted. By specifying a certain range of models, and by specifying *how* $\Omega(x)$ is to be interpreted at each of these models, this understanding provides a fair bit of semantic information about $\Omega(x)$.

Second, we can obtain *different* model-theoretic versions of $\Omega(x)$ simply by choosing different classes of models for our language and/or by specifying different ways of interpreting the individual components of $\Omega(x)$ at these models. So, for instance, we can give the sentence $\forall P \exists x P(x)$ different kinds of model-theoretic semantics by varying the rules for interpreting the "second-order" quantifier. If we insist that this quantifier range over all subsets of a domain, we get one kind of semantics; if we make it range over the *first-order definable* subsets of a domain, we get another; if we allow it to range over some *arbitrary* subset

[21] So, for instance, generic first-order semantics does not provide a universal domain for its quantifiers, but it does require that its quantifiers range over sets. Similarly, once we have a domain of quantification, first-order semantics does not tell us how to interpret particular predicates and constants, but it *does* require that this interpretation be correlated with the choice of domain for the quantifiers (i.e., constants must name *members* of this domain, and predicates must pick out *subsets* of the domain). Hence, although a generic first-order semantics leaves a great deal of room for varying interpretations, it also places substantial restrictions on those interpretations. See [15], [25], [53] or section 3.2 of this thesis for further discussion of this issue.

of the domain's powerset we get still a third. In this sense, then, model theory provides a whole collection of semantics for $\Omega(x)$, depending on how we choose to specify the relevant conceptions of "model" and/or "satisfaction."[22]

Finally, and most importantly, any model-theoretic semantics worth its salt will be *designed* to allow some variation in the models at which a particular sentence can be interpreted (and, indeed, at which a particular sentence can come out true). The point of model theory, after all, is to investigate the *interaction* between models and formulas. If we give *too specific* a semantics to our formulas—e.g., by fixing *everything* about the interpretation of our language and leaving *nothing* to vary as we move from model to model—then we render those formulas model-theoretically trivial.[23] Thus, the very nature of model theory requires that, in giving a particular formula a model-theoretic semantics, we allow that formula to be interpreted at a fairly wide (but interestingly limited) collection of models. It's simply built into the notion of semantics which underlies the study of model theory.[24]

For the purposes of this section, I will label the version of $\Omega(x)$ which gives this formula a generic first-order, model-theoretic semantics $\Omega_1(x)$. Once this semantics is in place, it immediately induces another version of $\Omega(x)$ (or, better, a whole family of versions of $\Omega(x)$) which results from interpreting $\Omega_1(x)$ *at* a particular model. Specifically, the quantifiers, which were allowed to range over an arbitrary set in $\Omega_1(x)$, are now fixed so as to range over the particular set which constitutes our model's domain. Similarly, the membership symbol, which was left (at least partially) undefined on the generic, first-order version of $\Omega(x)$, is now given a particular significance by the interpretation function of our model. Given a particular model $N$, I call the resulting interpretation of $\Omega(x)$, "$\Omega_N(x)$."

Before moving on, we should notice that the semantics of $\Omega_N(x)$ are in one way closer to the ordinary English semantics of $\Omega_E(x)$ than to the generic model-theoretic semantics from which it came. For one thing, this semantics gives determinate significance to *all* the components of $\Omega(x)$. For another, it gives a determinate truth-value to *all sentences* in the language of $\Omega(x)$. Like $\Omega_E(x)$, therefore, $\Omega_N(x)$ can be regarded as a "fully-interpreted" version of $\Omega(x)$.[25]

[22] Note that these are two kinds of "variability" in play here. One is the variability which comes when we move from model to model, *given* a generic semantics for $\Omega(x)$. The other comes when we make our initial choice of generic semantics—i.e., when we decide whether $\Omega(x)$ is a first-order formula, a second-order formula, or something else entirely. The paragraph above is concerned, solely, with the second of these types of variability.

[23] In particular, we should not be surprised to find that first-order sentences are typically satisfied by a whole variety of structurally different models. In designating a sentence "first-order," we say that it is to be evaluated at *these kinds* of models. And while *some* sentences—e.g., $\forall x \forall y\ x = y$—may do a good job at picking out the structure of their models, this cannot be the case for all sentences. If it were, then first-order semantics would lose most of its mathematical interest.

[24] Given this, philosophers should be more careful in their use of phrases like "fully-interpreted, first-order languages." From a model-theoretic perspective, the difference between first and second-order languages *depends* on those languages *lacking* a "full interpretation." There is, for instance, no interesting difference between fully-interpreted second-order arithmetic and a fully-interpreted first-order theory of numbers and sets of numbers. Only when we allow these two theories to take on *multiple* models—i.e., when we *remove* the "full interpretation—does the first-order/second-order distinction begin to do anything.

[25] Of course, exactly *how* $\Omega(x)$ gets interpreted depends on *which* model we chose to interpret it *at:* if $N \neq N'$, then $\Omega_N(x)$

All that being said, there are substantial *structural* differences between $\Omega_E(x)$ and $\Omega_N(x)$. Exploring these differences will be the principal focus of sections 1.3.2 and 1.3.3. For now, I conclude the present section by recalling the six items—two sentences of ordinary English and four versions of $\Omega(x)$—which have been introduced in this section: 1.) "x is uncountable," 2.) Cantor(x), 3.) $\Omega_E(x)$, 4.) $\Omega_P(x)$, 5.) $\Omega_1(x)$, and 6.) $\Omega_M(x)$ (where $M$ is as in argument (A) ). Understanding how these items contribute to argument (A) is the task of section 1.2.2.

### 1.2.2 The Heart of Skolem's Paradox

In this section, I discuss four issues concerning the relationship between argument (A) and the six items introduced in the last section. First, I argue that premise 2 in argument (A) is relatively plausible *if* "$\Omega(x)$" is interpreted as $\Omega_E(x)$. Second, I note that premise 3 amounts to the assertion that $\Omega_M(x)$ is true. Third, I use these first two points to highlight a key philosophical assumption upon which Skolem's Paradox rests—namely, the assumption that there is a deep relationship between the semantics of $\Omega_E(x)$ and those of $\Omega_M(x)$. Finally, I make a few brief remarks concerning the relevance of $\Omega_P(x)$ and $\Omega_1(x)$ to Skolem's Paradox.

To begin, consider the four versions of $\Omega(x)$ which were presented in the last section. Of these, $\Omega_E(x)$ is clearly the version which is most *naturally* related to the English phrase "$x$ is uncountable." Whereas the other versions all involve a fair bit of technical machinery—the introduction of a formal syntax in each case, and the introduction of a formal semantics in the case of $\Omega_1(x)$ and $\Omega_M(x)$—$\Omega_E(x)$ is essentially a sentence of ordinary English. Furthermore, the way $\Omega_E(x)$ is *derived* from "$x$ is uncountable—i.e., through the abbreviation of a sentence which is a direct explication of "x is uncountable"—makes it plausible to think that $\Omega_E(x)$ "says that" $x$ is uncountable.

To evaluate this thought more carefully, we need to ask ourselves whether that "says that" relationship is preserved through the following diagram—whether, that is, each of the three sentences in the diagram really "says" the same thing:

$$\text{"}x\text{ is uncountable"} \iff \text{Cantor(x)} \iff \Omega_E(x).$$

Start with the transition between "$x$ is uncountable" and Cantor($x$). Clearly there is a truth-functional equivalence between these two sentences. After all, Cantor(x) is simply a long-winded explication of "$x$ is uncountable," and any explication worth its salt preserve the truth-value of its explinandum. Nevertheless, I have a concern as to whether "$x$ is uncountable" and Cantor($x$) really "say" the same thing.

My concern here is this: there are many different ways to explicate the notion of uncountability. These explications differ in the ways they "code" basic set-theoretic concepts. So, for instance, some explications of "uncountability" use Kuratowski pairs to formulate the notion of "a bijection" while others use more

may be very different $\Omega_{N'}(x)$ (and may, indeed, have a different truth-value). This is simply to reiterate a point made in the last paragraph: the technique of specializing to a particular model provides a *family* of different versions of $\Omega(x)$. While these versions share certain family resemblances (induced by the fact that they all arise from first-order models for the language of set theory), they are still *different* interpretations of $\Omega(x)$.

cumbersome codings.[26] Similarly, some explications of uncountability emphasize the lack of a bijection between $x$ and $\omega$, while others emphasize the lack of a bijection between $x$ and some other set—say, the set of Zermelo numbers or the set of hereditarily finite sets.

Now, if all these different explications "said" the same thing, then there would be no problem in moving between "$x$ is uncountable" and Cantor($x$). However, I see no reason to think that these explications *do* say the same thing. Certainly it is possible for two such explications—$\text{Cantor}_1(x)$ and $\text{Cantor}_2(x)$—to fail to be logically equivalent. In such a case, it's hard to see why the two explications should be regarded as "saying" the same thing. As a result, it's hard to see why any one of these explications—e.g., Cantor(x)—should be singled out as *the explication* which "says the same thing" as "x is uncountable."[27]

Although I think this point is important for our understanding of set-theoretic English, I don't want to make too much of it here. For one thing, "$x$ is uncountable" and Cantor($x$) *are* truth-functionally equivalent (in every possible world, no less!). For another, a simple modification of argument (A) allows us to bypass my concern altogether: we just replace each instance of the term "countable" in argument (A) with some particular explication of countability—e.g., "it is not the case that Cantor(x)." The resulting argument raises the same problems as argument (A), and it does not involve any (explicit) use of the term "countable." In light of these considerations, I will proceed to grant—at least for the sake of argument—that "x is uncountable" and Cantor(x) really do "say the same thing."

With this concession in hand, the move to $\Omega_E(x)$ becomes unproblematic. Since $\Omega_E(x)$ is simply an abbreviation for Cantor(x), these two sentences have exactly the same semantics. Hence, the move between Cantor(x) and $\Omega_E(x)$ is trivial. If we combine this with the concession made in the last paragraph, then we are able to conclude that "$x$ is uncountable" and $\Omega_E(x)$ also say the same thing—i.e., that the "says that" relation is preserved through the three columns in the diagram above.

This, then, provides us with an acceptable reading of premise 2: if we stipulate that "$\Omega(x)$" in premise 2 is to be read as $\Omega_E(x)$, then premise 2 comes out true.[28] Further, it's fairly easy to see how premise 3 can also be formulated so as to tie into the material from the last section. Because the quantifiers in $\Omega_M(x)$ are taken to range over the domain of $M$ and because the membership symbol in $\Omega_M(x)$ is taken to mean $\in_M$, there is a perfect correspondence between the truth-conditions for $\Omega_M(x)$ and those for $M \models \Omega[x]$. In particular, for any element $m \in M$,

$$\Omega_M(m) \text{ is true } \iff M \models \Omega[m].$$

Hence, premise 3 turns out to be just another way of asserting $\Omega_M[\hat{m}]$.

[26] Kuratowski identifies the ordered pair $\langle a, b\rangle$ with the set $\{\{a\}, \{a, b\}\}$. We could just as easily identify it with the set $\{\{\{\{a\}\}, \{\{a, b\}\}\}\}$.

[27] Of course, we might argue that there is a *best* explication of "x is uncountable" and that this is the one which "says the same thing" as "x is uncountable." Since I will eventually grant—for the sake of argument—that Cantor(x) and "x is uncountable" *do* say the same thing, I will not pursue this suggestion here.

[28] This certainly works if "says that" is understood to mean nothing more than "is equivalent to." It may also work if "says that" is given a stronger reading, but this is far less clear (for the reasons mentioned in the main text).

At this point, we are ready to approach the philosophical heart of Skolem's Paradox: the assumption that there exists a connection between the semantics of $\Omega_E(x)$ and those $\Omega_M(x)$. For starters, we should note that the two conclusions we have already reached—i.e., that $\Omega_E(x)$ "says that" $x$ is uncountable, and that $\Omega_M(x)$ is equivalent to $M \models \Omega[x]$—allow us to reformulate the claim (†) in the following, somewhat more perspicuous, form:

$$\forall m \in M\, [\Omega_M(m) \implies \Omega_E(\{x \mid x \in_M m\})]. \tag{$\dagger'$}$$

This claim captures—in a purely truth-functional manner—the hypothetical connection between $\Omega_E(x)$ and $\Omega_M(x)$ which, I think, ultimately underlies Skolem's Paradox. In section 1.3, I discuss the truth of this claim. At present, I focus on the role it plays in the *formulation* of argument (A). There are three things to notice here.

*First*, ($\dagger'$) provides a necessary condition for argument (A) to be sound.[29] To see this, suppose that argument (A) is sound. Then the validity of the inference from 1–3 to 4 means that premises 1 and 2 must entail the conditional:

$$M \models \Omega[\hat{m}] \implies \text{“}\{x \mid x \in_M \hat{m}\} \text{ is uncountable.”}$$

But, since neither 1 nor 2 makes any mention of $\hat{m}$, and since the only distinguishing thing *about* $\hat{m}$ is the fact that $M \models \Omega[\hat{m}]$, premises 1 and 2 must actually entail the more general:

$$\forall m \in M\, [M \models \Omega[m] \implies \text{“}\{x \mid x \in_M m\} \text{ is uncountable”}].$$

As we have already seen, the antecedent of the conditional here is equivalent to $\Omega_M(m)$ and the consequent is equivalent to $\Omega_E(\{x \mid x \in_M m\})$. Substituting in accordance with these equivalences, therefore, we obtain ($\dagger'$). Finally, since argument (A) has been assumed to be sound, premises 1 and 2 must be true. So, ($\dagger'$) must be true as well.

It is important to notice that the argument above does not depend on any particular interpretation of the "$\Omega(x)$" in premise 2. Nor does it depend on the details of premise 1. Instead, it depends on—and only on—the two facts discussed earlier in this section: that $\Omega_E(x)$ says the same as "$x$ is uncountable" and that $\Omega_M(x)$ is equivalent to "$M \models \Omega[x]$." Given these two facts, *any* argument which moves from premise 3 of argument (A) to line 4 of argument (A) must also presuppose—or at least imply—($\dagger'$).

*Second*, if we focus on the version of argument (A) which interprets the "$\Omega(x)$" in premise 2 as $\Omega_E(x)$, then ($\dagger'$) provides a sufficient condition for (A) to be sound. On the one hand, and as we have already seen, premises 1, 3 and 5 in argument (A) are clearly true; if we use $\Omega_E(x)$ as the interpretation of "$\Omega(x)$," then premise 2 is true as well. On the other hand, we know that the inference from 1, 3 and 5 to 6 is valid; in the presence of ($\dagger'$), the inference from 1–3 to 4 is also valid. Thus, ($\dagger'$) is a sufficient condition for the soundness of argument (A).

[29] Philosophers typically use the word "sound" to mean that an an argument is valid *and* that it has true premises. Mathematicians typically use "sound" as a synonym for "valid." Here, and throughout this thesis, I will follow philosophical usage and assume that sound arguments have (only) true premises.

*Third*, ($\dagger'$) is a relatively *natural* assumption to make about $\Omega_E(x)$ and $\Omega_M(x)$. Because these two sentences have the same syntactic structure, it is easy to assume that they have the same semantics as well. Hence, it is easy to assume that we can infer $\Omega_E(x)$ from $\Omega_M(x)$ in the course of an argument. More importantly, even if we are inclined to *reject* the inference from $\Omega_M(x)$ to $\Omega_E(x)$, it is often easy to miss the fact that this inference is *involved* in a given presentation of Skolem's Paradox. If Skolem's Paradox is presented rapidly, or is presented so as to slur over the distinctions made in 1.2.1—say, by using a single phrase to mean $\Omega_E(x)$ at some points and $\Omega_M(x)$ at others—then it is easy to lose sight of the fact that a transition between $\Omega_M(x)$ and $\Omega_E(x)$ even takes place.[30]

These, then, are the reasons I regard ($\dagger'$) as the crucial philosophical assumption on which Skolem's Paradox ultimately rests. From a logical standpoint, ($\dagger'$) is presupposed by *every* version of Skolem's Paradox. In addition, ($\dagger'$) is sufficient to make one version of Skolem's Paradox go through: if ($\dagger'$) is true, then the version of argument (A) which uses $\Omega_E(x)$ in premise 2 comes out sound. Finally, ($\dagger'$) is an *natural* assumption: whether true or not, ($\dagger'$) captures much of what makes Skolem's Paradox feel paradoxical in the first place.

Before examining the truth of ($\dagger'$), I want to conclude this section by making two remarks about $\Omega_P(x)$ and $\Omega_1(x)$. First, $\Omega_P(x)$ and $\Omega_1(x)$ are essentially irrelevant for the purpose of understanding premise 2. Since $\Omega_P(x)$ has no semantics, it cannot "say" anything at all. And, although $\Omega_1(x)$ *does* have a semantics, this semantics is too indeterminate for $\Omega_1(x)$ to say that "$x$ is uncountable." Unless $\Omega_1(x)$ is interpreted *at* a specific model—i.e., unless it is turned into some $\Omega_N(x)$—it contains too many undefined (or incompletely defined) symbols to take on a specific meaning. Therefore, neither of these formulas is an acceptable candidate for use in premise 2.

Second, although $\Omega_P(x)$ and $\Omega_1(x)$ are closely related to $\Omega_M(x)$—e.g., in that they all share the same formal syntax and in that $\Omega_1(x)$ shares *some of* the semantics of $\Omega_M(x)$—they do not provide much help in understanding ($\dagger'$). In particular, we can't justify ($\dagger'$) by arguing that the "says that" relation—or even some weak truth-functional analog of this relation—is preserved *through* one of the following diagrams:

$$\begin{tikzcd}
\Omega_M(x) \arrow[r, Rightarrow] & \Omega_P(x) \arrow[r, Rightarrow] & \Omega_E(x) \\
\Omega_M(x) \arrow[r, Rightarrow] & \Omega_1(x) \arrow[r, Rightarrow] & \Omega_E(x) \\
(1) & (2) & (3)
\end{tikzcd}$$

In each case, the transition between column (1) and column (2) tries to connect a sentence having a specific semantics—and a particular truth value—to a sentence which has (at best) an indefinite semantics—and,

[30] In my view, unconscious slides between $\Omega_E(x)$ and $\Omega_M(x)$ are the most important reason for philosophers taking Skolem's Paradox seriously. To be sure, some philosophers have defended this slide explicitly; their arguments will be examined in chapters 3 and 4. In general, however, I think that our initial feeling that there is something to Skolem's Paradox—that Skolem's Paradox really *is* a paradox—stems less from explicit arguments than from implicit slides between $\Omega_M(x)$ and $\Omega_E(x)$ (slides which may be facilitated by an equivocal use of phrases like "x is uncountable" and "Omega(x).").

in consequence, no truth value at all. Similarly, the transition between column (2) and column (3) tries to connect a sentence with no truth value to a sentence with a specific truth value. Hence, even though $\Omega_M(x)$ requires $\Omega_P(x)$ and $\Omega_1(x)$ for its formulation, we should not expect its relations with $\Omega_E(x)$ to work *through* $\Omega_P(x)$ and $\Omega_1(x)$—i.e., we should not expect these sentences to function as *intermediaries* between $\Omega_M(x)$ and $\Omega_E(x)$. Thus, whatever interest $\Omega_P(x)$ and $\Omega_1(x)$ have for mathematicians, they do very little towards helping us understand argument (A).

At this point, I think we have obtained a fairly detailed understanding of the six items introduced in section 1.2.1, along with the roles these items play in argument (A). Although $\Omega_P(x)$ and $\Omega_1(x)$ help us to understand $\Omega_M(x)$, they do not play an independent role in argument (A). In contrast, $\Omega_E(x)$ and $\Omega_M(x)$ are quite crucial to this argument. $\Omega_E(x)$ provides the most plausible reading of "$\Omega(x)$" in premise 2; $\Omega_M(\hat{m})$ provides an alternate formulation of premise 3. Together, they allow us to formulate the key assumption which underlies argument (A)—i.e., the claim $(\dagger')$. In section 1.3 (esp. 1.3.2) I turn, at last, to the assessment of this assumption.

## 1.3 Solving Skolem's Paradox

This section has two goals: 1.) to evaluate the truth of $(\dagger')$ and 2.) to provide a "solution" to the version of Skolem's Paradox which we have been discussing. The bulk of this work takes place in 1.3.1–1.3.3. In 1.3.1, I do two things. First, I argue that any plausibility that $(\dagger')$ may have depends on the fact that $M$ is a model for ZFC. Second, I argue that $(\dagger')$ is false and that this falsity should be obvious once we understand the role $(\dagger')$ plays in Skolem's Paradox. Once this argument is completed, I turn in 1.3.2 and 1.3.3 to explain *why* $(\dagger')$ is false: i.e., *where* the semantics of $\Omega_M(x)$ and $\Omega_E(x)$ differ and *how* this difference leads to the failure of $(\dagger')$. Finally, in section 1.3.4, I step back to discuss some more general issues concerning the "solution" to Skolem's Paradox which this chapter has provided.

### 1.3.1 The Solution

Let's begin with $(\dagger')$. Insofar as $(\dagger')$ depends on the model $M$, it is reasonable to wonder just which *features* of $M$ give $(\dagger')$ its intuitive plausibility. On my reading, the key fact about $M$ which makes $(\dagger')$ seem plausible is the fact that $M \models$ ZFC. To reinforce this thought, let me make a few remarks concerning the generalization of $(\dagger')$ to arbitrary models.

For convenience, I use $(\dagger'_G)$ to represent the following, rather broadminded, generalization of $(\dagger')$:

$$\forall N \forall n \in N\, [\Omega_N(n) \implies \Omega_E(\{x \mid x \in_N n\})]. \qquad (\dagger'_G)$$

The first thing to notice here is that we lack even *prima facie* reasons for accepting this broad generalization of $(\dagger')$. After all, there are many models for the language of set theory which contain objects other than sets, and there are some models which contain no sets at all.[31] In such cases, it's hard to see why these models should be regarded as having anything to do with set theory, *unless* they happen satisfy some set-theoretic axioms—say, some significant fragments of ZFC. In consequence, it's hard to see why we should accept any instance of ($\dagger'_G$) which relates to such models—i.e., any instantiation of ($\dagger'_G$) to an $N$ such that $N \not\models$ ZFC.

To reinforce this point, we should notice *just how badly* models for the language of set theory can fail to satisfy ZFC, while nevertheless satisfying formulas like $\Omega[m]$. Consider, for instance, the model whose domain consists of the numbers 1–10 and which interprets "$\in$" by:

$$n \in m \Longleftrightarrow n \leq 5 \text{ and } 5 < m \leq 10.$$

In this model, all numbers greater than 5 satisfy $\Omega(x)$, although the model itself has no connection with set theory and fails to satisfy even the axiom of extensionality.[32] For that matter, if we let $\Psi(y)$ be the formula which codes "$y = \omega$," then *any* model which satisfies "$\neg\exists y\, \Psi(y)$" will also satisfy "$\forall x\, \Omega(x)$."[33]

These examples show how badly a model can fail to satisfy ZFC while still satisfying formulas like $\Omega[m]$. The first example also proves (outright) that ($\dagger'_G$) is false. For suppose that ($\dagger'_G$) were true. Then, instantiating $N$ to the model above and instantiating $n$ to 6, we would get the conditional:

$$\Omega_N(6) \implies \Omega_E(\{1, 2, 3, 4, 5\}).$$

But, since $\Omega_N(6)$ is true, this conditional entails that $\{1, 2, 3, 4, 5\}$ is uncountable. This is absurd.

Finally, note that, even if ($\dagger'_G$) *were* true, this fact would be of limited use to proponents of Skolem's Paradox. After all, ($\dagger'_G$) is such a strong generalization of ($\dagger'$) that it facilitates variants of argument (A) which 1.) reach the same conclusions as argument (A) but 2.) make *no* use of the Löwenheim-Skolem theorems.[34] Hence, far from assisting in the formulation of Skolem's Paradox, ($\dagger'_G$) actually *trivializes* this

[31] So, for instance, we have already seen a model $N$ such that $N \models$ "Gandalf $\in$ Katie" where Gandalf and Katie are two ordinary housecats. See 10.

[32] With respect to the axiom of extensionality, note that all of the numbers $n \leq 5$ have exactly the same "members," as do all of the numbers $m > 5$. With respect to the satisfaction of $\Omega(x)$, note that this formula has the overall form:

$$\Omega(x) \equiv_{df} \neg\exists f\ [\text{"}f \text{ is a function"} \ \&\ \text{Domain}(f) = \omega \ \&\ \text{Range}(f) = x].$$

Here, the phrases "$x$ is a function," "Range($f$)," and "Domain($f$)" are themselves mere abbreviations for further (rather complicated) formulas. For our purposes, the important thing to notice is that the formulas "$f$ is a function" and "Range($f$) $= x$" together entail that every member of $x$ is also a member of a member of a member of $f$. Hence, since the interpretation of "$\in$" in our model does not allow membership chains containing more than two elements, no $f$ of the type forbidden by, e.g., $\Omega[6]$ lives in our model. Hence, the model satisfies $\Omega[6]$ (and $\Omega[7]$, and $\Omega[8]$, etc.).

It is worth noting that this model *also* satisfies $\Omega[n]$ for $n \leq 5$, though unpacking the relevant definitions is more time-consuming in these cases and depends on a particular definition of $\omega$. The basic idea is that discussed in the next footnote.

[33] Again, this is a simple consequence of the definition of $\Omega(x)$. To see this, simply note that:

$$\neg\exists y\, \Psi(y) \vdash \neg\exists f\ [\cdots \ \&\ \exists y\ (y = \text{Domain}(f) \ \&\ \Psi(y)) \ \&\ \cdots]$$

for *any* possible values of "$\cdots$" ( including those relevant to $\Omega(x)$).

[34] To obtain one such argument, we simply replace $M$ in (A) with the model discussed in the last paragraph and then replace

paradox. Given ($\dagger'_G$), we can derive the same conclusions as argument (A) without mentioning the fact that ZFC has countable models.

These, then, give three reasons for eschewing ($\dagger'_G$). First, ($\dagger'_G$) not a very plausible principle: given that some models contain no genuine sets and fail to satisfy even the most basic set-theoretic axioms, there's no reason to think that ($\dagger'_G$) should be true of those models. Second, easy counterexamples show that ($\dagger'_G$) is actually false. Third, even if ($\dagger'_G$) *were* true, this would only serve to trivialize—rather than facilitate—our overall formulation of Skolem's Paradox.

Together, these three points reinforce the thought that $M$'s satisfaction of ZFC plays an important role in making ($\dagger'$) intuitively plausible. Unfortunately, the mere fact that $M \models \text{ZFC}$ is not enough to make ($\dagger'$) *true*. To see this, we should begin by considering the following weakening of ($\dagger'_G$):

$$\forall N\,[N \models \text{ZFC} \Longrightarrow \forall n \in N\,[\Omega_N(n) \implies \Omega_E(\{x \mid x \in_N n\})]]. \tag{$\dagger'_{\text{ZFC}}$}$$

To see that ($\dagger'_{\text{ZFC}}$) is false, we need only let $N$ be a countable model of ZFC, and let $\hat{n} \in N$ be such that $N \models \Omega[\hat{n}]$. Then, employing the equivalences introduced in 1.2.2, we have that $N \models \text{ZFC}$, that $\Omega_N(\hat{n})$ is true, and that $\Omega_E(\{n \mid n \in_N \hat{n}\})$ is false. Therefore, ($\dagger'_{\text{ZFC}}$) is false as well.

This example shows that merely adding a condition about the satisfaction of ZFC to ($\dagger'$) will not make ($\dagger'$) true in any generalizable way. More importantly, the example actually shows that ($\dagger'$) is false—i.e., false even in the *ungeneralized* form involving only the particular model $M$. After all, the way that $N$ and $\hat{n}$ were chosen in the example above runs *exactly* parallel to the way that $M$ and $\hat{m}$ were chosen at the beginning of this chapter. Hence, everything said about $N$ and $\hat{n}$ must be true of $M$ and $\hat{m}$ as well. In particular, $\Omega_M(\hat{m})$ is true and $\Omega_E(\{m \mid m \in \hat{m}\})$ is false. So ($\dagger'$) is false as well.

There are two things we should notice about this argument. First, although I have taken the time to give an explicit argument against ($\dagger'$), the actual falsity of ($\dagger'$) should have been obvious from the role ($\dagger'$) plays in Skolem's Paradox. As we saw in the last section, ($\dagger'$) is a sufficient condition for one version of argument (A) to be sound. Since this version of argument (A) leads to a contradiction, the claim that it is sound must be rejected. Hence, ($\dagger'$) must be rejected as well.

Second, the fact that ($\dagger'$) is false renders *all versions* of argument (A) unsound. As we saw in the last section, ($\dagger'$) is a necessary condition for any version of argument (A) to work; so, the falsity of ($\dagger'$) entails that no version of argument (A) can possibly work. Even though our explicit argument against ($\dagger'$) depends on the role ($\dagger'$) plays in *one particular* version of argument (A)—i.e., the version which ($\dagger'$) renders sound—the ultimate failure of ($\dagger'$) undercuts *every* version of argument (A).

This, then, gives us a solution, of sorts, to the version of Skolem's Paradox which we have been discussing. We have isolated a key principle on which argument (A) depends, and we have shown that this principle provides a necessary condition for any version of argument (A) to be sound. Further, we have given two

$\hat{m}$ with 6. Next, we modify premise 1 of (A) to reflect our changed model. All of the premises in this new argument are clearly true, and the argument is valid if ($\dagger'_G$) is true (indeed, just in case ($\dagger'_G$) is true).

arguments—one direct and one indirect—which prove that this principle is false. As a result, we have an explanation as to *where* Skolem's Paradox actually fails.

All that being said, this "solution" is clearly missing something. Although we know that $(\dagger')$ is the place where argument (A) goes wrong, we don't yet know where $(\dagger')$ itself goes wrong. Put another way: although we know *that* $(\dagger')$ is false, we have yet to discover *why* $(\dagger')$ is false. What we really need, therefore, is an analysis of the semantic differences between $\Omega_E(x)$ and $\Omega_M(x)$ which explains why the former does not entail the latter (or, at the very least, why the two are sufficiently different that we should not be *surprised* when the former fails to entail the latter with respect to a particular model $M$).

The need for this analysis can be underscored by noting a way in which this "solution" may seem to miss the point of Skolem's Paradox. The proponent of Skolem's Paradox is sufficiently convinced that there exists a relationship between $\Omega_M(x)$ and $\Omega_E(x)$—a relationship strong enough to ground $(\dagger')$—that he is willing to reconstrue set theory in light of this conviction. In particular, the proponent of Skolem's Paradox claims that set theory, when taken at face value, *just is* contradictory; he then turns to notions like *relativity* and *perspective* to ease the philosophical sting of this contradiction.

Given this, it is highly unlikely that the proponent of Skolem's Paradox will be persuaded by the kind of *modus tollens* argument which I have just given (or even the direct counterexample which precedes it). The proponent already *knows* that his assumptions lead to a contradiction—that, after all, is the whole *point* of Skolem's Paradox. By itself, however, this contradiction is not enough to make him abandon his assumptions. Hence, unless the arguments above are supplemented by a detailed analysis of *why* $(\dagger')$ fails—i.e., of *where* the semantics of $\Omega_M(x)$ and $\Omega_E(x)$ differ and of *how* this difference leads to the failure of $(\dagger')$—the proponent of Skolem's Paradox is unlikely to find them very persuasive.

### 1.3.2 Why $(\dagger')$ Fails I

How, then, *do* $\Omega_M(x)$ and $\Omega_E(x)$ differ? I think there are two different, and differently significant, answers to this question. First, to the extent that the semantics of ordinary English involves things like *meanings* or *senses*—i.e., as opposed to being simply referential—this induces a difference between the semantics of $\Omega_E(x)$ and $\Omega_M(x)$. When a formula like $\Omega_1(x)$ is interpreted at a model, the interpretation does not work *through* any intentional apparatus—e.g., any meanings or senses. Instead, the interpretation function *directly* associates the symbol "$\in$" to some particular relation on the domain of $M$.

Given this, questions about "what '$\in$' means in $M$" are somewhat misleading. In model theory, individual bits of language don't *mean* anything at all; they simply sit in the domain of an interpretation function which takes pieces of a model for its range. So, to the extent that sentences like $\Omega_E(x)$ *do* have genuine meanings—i.e., to the extent that ordinary mathematical English involves meanings—there will be semantic differences between $\Omega_E(x)$ and $\Omega_M(x)$.[35]

[35] In practice, *these particular* semantic differences won't help to explain Skolem's Paradox. Hence, readers who are only interested in Skolem's Paradox may ignore the remainder of this section and skip directly to 1.3.3.

To highlight some of these differences, it's useful to examine the behavior of ordinary English semantics and model-theoretic semantics in certain modal contexts. Suppose that we are interested in cats (rather than sets) and that we are considering the sentence: "Gandalf and Katie are the only two cats who live in Tim's house." Following our initial formulation of $\Omega_E(x)$, we could abbreviate this sentence by the following:[36]

$$C(g) \wedge C(k) \wedge g \neq k \wedge H(t,g) \wedge H(t,k) \wedge \neg\exists x[C(x) \wedge H(t,x) \wedge x \neq g \wedge x \neq k]$$

Let's call this abbreviation "$\Psi_E(g,k,t)$." Despite its formal appearance, this is a sentence of ordinary English, and its relationship to the world is governed by ordinary English semantics.

In contrast, suppose that $N$ is a model for the formal language containing the symbols $C, H, g, k,$ and $t$. The domain of $N$ is just the three element set $\{\text{Gandalf}, \text{Katie}, \text{Tim}\}$, and the interpretation function for $N$ works in the obvious way. Following the construction of $\Omega_M(x)$ in section 1.2.1, we find that $N$ induces a sentence which is syntactically similar to $\Psi_E(g,k,t)$, but which gets its semantics via an interpretation *in* $N$. Let's call this sentence $\Psi_N(g,k,t)$.

Now, although $\Psi_E(g,k,t)$ and $\Psi_N(g,k,t)$ happen to have the same truth-value in *this* possible world, we can imagine possible worlds where these truth values differ. Suppose, for instance, that I were to bring some new kittens into my household. In that case, $\Psi_E(g,k,t)$ would become false—since there would be extra cats in the house—but $\Psi_M(g,k,t)$ would continue to be true—since there would be *no* extra elements in $\{\text{Gandalf}, \text{Katie}, \text{Tim}\}$. Similarly, suppose that I were to tire of my cats and send them to the pound. Once again, $\Psi_E(g,k,t)$ would become false, but $\Psi_N(g,k,t)$ would remain true as ever.

These examples show that $\Psi_N(g,k,t)$ has a kind of modal stability which $\Psi_E(g,k,t)$ lacks. Because some of the terms in $\Psi_E(g,k,t)$ acquire their reference *through* their senses, these references change as we move from world to world.[37] In contrast, terms in $\Psi_N(g,k,t)$ have their "reference" fixed by the interpretation function of $N$; as this function does not change from world to world, the truth-value of $\Psi_N(g,k,t)$ does not change either. This discrepancy between the modal status of $\Psi_E(g,k,t)$ and $\Psi_N(g,k,t)$ highlights the difference between the ordinary English semantics of $\Psi_E(g,k,t)$ and the model-theoretic semantics of $\Psi_N(g,k,t)$.[38]

Of course, this example does not directly involve the sentences $\Omega_E(x)$ and $\Omega_M(x)$, and there are reasons to think that no similar examples can be constructed for these sentences. After all, $\Omega_E(\hat{m})$ is a *mathematical*

[36] Actually, this way of putting things skips a step. Ideally, we should first paraphrase our sentence into a more "austere" formulation—say, "Cats(g, k, t)." We would then use this paraphrase as the basis for abbreviation. As this detour doesn't add much to the current exposition, I omit the details here.

[37] In saying this, I do not mean to suggest that senses can be *identified* with functions between possible worlds and truth-values—i.e., with so-called "Carnap Propositions." It is enough for my purposes that senses help to *explain* the modal behavior of sentences like $\Psi_E(g,k,t)$, and that this modal behaviour, in turn, helps to distinguish sentences like $\Psi_E(g,k,t)$ from sentences like $\Psi_N(g,k,t)$.

[38] This distinction can also be played out in the other direction. Suppose that my cat Gandalf were to die tomorrow. Then it is plausible to think that $\Psi_E(g,k,t)$ would still be meaningful and would still have a truth value. In contrast, $\Psi_N(g,k,t)$ would almost certainly become nonsense, since the set $\{\text{Gandalf}, \text{Katie}, \text{Tim}\}$ and (hence) the model $N$ would no longer exist. I am grateful to Tony Martin for bringing this kind of example to my attention.

sentence, and mathematical sentences possess the same kind of modal stability as $\Omega_M(\hat{m})$—i.e., they are either necessarily true or necessarily false. As a result, the phenomenon discussed in the last two paragraphs cannot be used (directly) to distinguish between $\Omega_E(x)$ and $\Omega_M(x)$.

Nevertheless, we should remember that the modal distinction between $\Psi_E(g,k,t)$ and $\Psi_N(g,k,t)$ served only to *highlight* a deeper difference between the semantics of these two sentences. This deeper difference stems purely from the fact that $\Psi_E(g,k,t)$ is a sentence of ordinary English while $\Psi_N(g,k,t)$ is a formal sentence with a model-theoretic semantics. Hence, even if no modal difference between $\Omega_E(x)$ and $\Omega_M(x)$ can be discerned, this deeper difference should still carry over. As a sentence of ordinary English, $\Omega_E(x)$ has a semantics which works *through* the meanings of phrases like "is a set," "is a member of," etc. In contrast, $\Omega_M(x)$ has a model-theoretic semantics in which no meanings are involved.

Leaving modality aside, it's possible to give a *direct* argument for the kind of semantic difference at issue here. To see this, begin with the sentence Cantor(x), and substitute the phrase "__ is a member of a member of the singleton of __, if there exist at least two measurable cardinals" for every occurrence of "__ is a member of __". Call the resulting sentence Cantor$'(x)$. Next, abbreviate Cantor(x) in the manner described in section 1.2.1, and abbreviate Cantor$'(x)$ so as to use the symbol "$\in$" to represent the phrase mentioned above.[39] Call the resulting abbreviations $\Omega_E(x)$ and $\Omega'_E(x)$ respectively.[40]

With respect to these formulas, we should notice four things. First, because of the way they were constructed, $\Omega_E(x)$ and $\Omega'_E(x)$ have *exactly* the same syntax. Second, although Cantor(x) and Cantor$'(x)$ are true for the same values of $x$, they almost certainly have different *meanings* (since, after all, the one involves the notion of measurable cardinals, and the other does not). Third, since $\Omega_E(x)$ and $\Omega'_E(x)$ are mere abbreviations for Cantor(x) and Cantor$'(x)$, the former have the same semantics as the latter; in particular, $\Omega_E(x)$ and $\Omega'_E(x)$ also have different meanings. Finally, and as a consequence of the first three points, *meanings* do play an important role in the semantics of sentences like $\Omega_E(x)$: even when two variants of this sentence look *exactly* alike and come out true under *exactly* the same circumstances, we can still distinguish them in virtue of their meanings.

In contrast, let's turn to model theory. Let $X$ be some transitive set and let $N$ and $N'$ be models for the language of set theory such that:

- Domain$(N) = X$ and Domain$(N') = X$.
- $\in_N = \{\langle x,y\rangle \mid x,y \in X \text{ and } x \in y\}$.
- $\in_{N'} = \{\langle x,y\rangle \mid x,y \in X \text{ and } \exists z\, [x \in z \in \{y\}] \text{ and there exist at least two measurable cardinals }\}$.

[39] Doing this right requires a little care. It is important that we *first* abbreviate the above-mentioned phrase and *then* abbreviate "not," "if ... then," etc. If we try to abbreviate the other items first, then we will inadvertently eliminate most instances of "__ is a member of a member of the singleton of __, if there exist at least two measurable cardinals."

[40] Throughout the following argument, I assume that (at least two) measurable cardinals exist. If this assumption is mistaken, then we can simply rephrase the example using "if there are fewer than two measurable cardinals" in place of "if there are at least two measurable cardinals." None of the content of the example depends on which way we go here.

The first thing to notice here is that $N$ and $N'$ are really the same model. The second is that, because $N$ and $N'$ are the same model, $\Omega_N(x)$ and $\Omega_{N'}(x)$ are really the same formula. As a result, there can be no difference between the semantics of $\Omega_N(x)$ and $\Omega_{N'}(x)$.

This example shows that model-theoretic semantics cannot "get at" the difference in meaning between "is a member of" and "is a member of a member of the singleton of ..." Even when these phrases are used to *define* the interpretation of "$\in$," the differences in meaning between the phrases don't effect the ultimate semantics of $\in_N$. Since the semantics of "$\in$" are totally fixed by $N$'s interpretation function—i.e., by the mere *set* $\{\langle \in, \{\langle x, y\rangle \mid x, y \in X \text{ and } x \in y\}\rangle\}$—distinctions in the way this set is *described* do not, ultimately, effect these semantics.

This, then, gives us our first difference between the semantics of $\Omega_E(x)$ and $\Omega_M(x)$. The semantics of $\Omega_E(x)$ are those of an ordinary English sentence. They work through some form of meaning, they have special behavior in modal contexts, and they allow fine-grained distinctions between truth-functionally equivalent sentences (e.g., $\Omega_E(x)$ and $\Omega'_E(x)$). In contrast, the semantics of $\Omega_M(x)$ are those of first-order model theory. They are purely extensional, they have rather different behavior in modal contexts, and they cannot make fine-grained distinctions between truth-functionally equivalent sentences.[41]

### 1.3.3 Why ($\dagger'$) Fails II

Leaving these intentional issues aside, let's examine some of the purely extensional (or truth-functional) differences in the ways the semantics of $\Omega_E(x)$ and those of $\Omega_M(x)$ interpret the specific symbols used in these two sentences. To begin, the semantics of $\Omega_E(x)$ interpret the symbol "$\in$" so that

E-$\in$: "$x \in y$" is true iff $y$ is a set and $x$ is a member of $y$,

while the semantics of $\Omega_M(x)$ interpret "$\in$" so that

M-$\in$: "$x \in y$" is true iff $\langle x, y\rangle$ is a member of $i_M(\in)$,

where $i_M$ is the interpretation function for $M$. As we have already seen, however, there is no reason to think that these two interpretations of "$\in$" are coextensive. This is clearest when the elements of $M$ are not even *candidates* for entering into the the ordinary membership relation. So, for instance, there exist models in which the relation $\in_M$ applies between ordinary housecats (cf. p. 10). Similarly, there are models whose

[41] The key thing to notice here is that this difference in semantics comes into play *even when* there is no referential difference between English semantics and model-theoretic semantics. This was certainly the case in my earlier example about cats, and I think it's true in the case of set theory as well.

Suppose, for instance, that we abuse Gödel a moment and treat $(V, \in)$ as a *model* for the language of set theory. In this case, there would be no truth-functional difference between $\Omega_E(x)$ and $\Omega_{(V,\in)}(x)$; nor would there be a difference in modal stability. Nevertheless, I think that $\Omega_E(x)$ involves the meanings of phrases like "is a set" and "is a member of" while $\Omega_{(V,\in)}(x)$ does not. After all, on the assumption that $\Omega_{(V,\in)}(x)$ *did* involve meanings, what would determine whether these meanings lined up with those of $\Omega_E(x)$ or those of $\Omega'_E(x)$? If they lined up with $\Omega_E(x)$, would there be *another* model which lined up with $\Omega'_E(x)$–say, $(V, \in)'$? What would be the model-theoretic difference between $(V, \in)'$ and $(V, \in)$?

domains contain no sets at all. In these cases, its quite clear that the semantics of $\Omega_M(x)$ and $\Omega_E(x)$ are interpreting the symbol "$\in$" in very different ways.

Even if a model *does* contain sets (and even *only sets*), there is no guarantee that the model's interpretation of "$\in$" agrees with the ordinary English interpretation of this symbol. To illustrate this point, let $N$ be an arbitrary model for the language of set theory, let $X$ be the collection of singletons of members of $N$ (e.g., $X = \{\{n\} \mid n \in N\}$), let $Y$ be the collection of doubletons of members of $X$ (e.g., $Y = \{\{x_1, x_2\} \mid x_1, x_2 \in X \text{ and } x_1 \neq x_2\}$), and let $A \subset Y$ such that $|A| = |N|$. Using a trick from footnote 6, we can find a model $N'$ such that 1.) $N'$ is isomorphic to $N$ and 2.) $N'$ has $A$ as its domain. In $N'$, therefore, all of the objects are genuine sets, but there is almost *no* agreement with ordinary English concerning the interpretation of "$\in$." In particular, there are many sets $n_1, n_2 \in N'$ such that $N' \models n_1 \in n_2$, but there are *no* sets $n_1, n_2 \in N'$ such that $n_1 \in n_2$.[42]

These examples show how the semantics of $\Omega_E(x)$ and those of $\Omega_M(x)$ sometimes disagree about atomic formulas. When we move to more complicated formulas, we find further disagreements. In particular, the semantics of $\Omega_E(x)$ interpret the expression "$\exists\, x$" as synonymous with the phrase "there is a set $x$, such that" (since, after all, the former is nothing more than an *abbreviation* for the latter). In contrast, the semantics of $\Omega_M(x)$ interpret the expression "$\exists\, x$" via the recursion clause:

$$\exists. \qquad M \models \exists\, x \Phi(x) \Longleftrightarrow \text{there exists an } m \in M \text{ such that } M \models \Phi[m].$$

In effect, this amounts to identifying the expression "$\exists\, x$" with the phrase "there is an element $x \in M$, such that." Since the domain of $M$ cannot be identical with the set-theoretic universe (as $M$ is, after all, a merely *countable* model), this introduces a second difference between the semantics of $\Omega_E(x)$ and those of $\Omega_M(x)$.

To close out this point, there are three things I want to notice. First, these two differences in the interpretation of "$\in$" and "$\exists\, x$" may lead to *many* differences between $\Omega_E(x)$ and $\Omega_M(x)$. Since each version of $\Omega(x)$ contains several thousand instances of "$\in$" and "$\exists\, x$," there are many places where the semantics of $\Omega_E(x)$ and $\Omega_M(x)$ can differ. Our two basic differences, therefore, have the potential to ramify into deeper—and more systematic—differences between the semantics of $\Omega_E(x)$ and those of $\Omega_M(x)$.

Second, these semantic differences exist *even when* $\Omega_E(x)$ and $\Omega_M(x)$ agree about some particular element of $M$—i.e., when $\Omega_E(m)$ and $\Omega_M(m)$ are both true (or false). If we check carefully, we often find that these sentences are true for structurally different reasons. The membership relations which make $\Omega_E(m)$ true may have nothing to do with the instances of $m_1 \in_M m_2$ which make $\Omega_M(m)$ true; the particular sets which make "there exists a set $x$ such that ..." true may be different from the elements of $M$ which make "there exists an $m \in M$, such that ..." true. As a result, occasional agreements between $\Omega_E(m)$ and $\Omega_M(m)$ may be little more than happy accidents.

Finally, and most importantly, the differences between the semantics of $\Omega_E(x)$ and $\Omega_M(x)$ should lead us to be unsurprised by the absence of a strong relationship between the truth values of these sentences. Since

[42] This latter claim follows from the fact that every element of $A$ is a doubleton and that every element of $A$ *contains* only singletons. Hence, there can be no two elements $a_1, a_2 \in A$, such that $a_1 \in a_2$.

many of the corresponding parts of $\Omega_E(x)$ and $\Omega_M(x)$ have different semantics—and semantics which differ in ways that effect the truth values of $\Omega_E(x)$ and $\Omega_M(x)$—there is no reason to expect the overall sentences to have any real relationship. As noted in the last paragraph, the differences between $\Omega_E(x)$ and $\Omega_M(x)$ are sufficiently severe that, if these sentences *were to* agree on all members of $M$'s domain, this *agreement* would be the fact in need of explanation. Where such agreement is lacking, we should regard this fact as ordinary and unsurprising.

This, then, gives a fairly complete solution to the version of Skolem's Paradox which we have been examining. In section 1.2.2, I showed that the paradox rests on the assumption that there exists a systematic relationship between the semantics of $\Omega_E(x)$ and $\Omega_M(x)$—i.e., the relationship captured by ($\dagger'$). In section 1.3.1, I showed that this assumption is clearly false. Finally, in sections 1.3.2 and 1.3.3, I have examined the details of $\Omega_E(x)$ and $\Omega_M(x)$ and have used this examination to explain why the falsity of ($\dagger'$) is neither surprising nor paradoxical.

### 1.3.4 Four Remarks

Before I conclude this section, I want to make four remarks concerning the solution to Skolem's Paradox which I have just given. First, the philosophical upshot of this solution is that Skolem's Paradox rests on an equivocation. To make premise 2 in argument (A) plausible, we need to read "$\Omega(x)$" as $\Omega_E(x)$. In contrast, premise 3 clearly involves $\Omega_M(x)$. By using "$\Omega(x)$" in an equivocal manner, we can assimilate $\Omega_M(x)$ to $\Omega_E(x)$ and thereby make argument (A) psychologically plausible. However, $\Omega_M(x)$ and $\Omega_E(x)$ are radically different sentences—so different that even ($\dagger'$) fails to hold. Hence, no argument which equivocates between these two sentences can possibly be valid.

Second, this solution to Skolem's Paradox is clearly tailored to the relatively simple formulation of that paradox that was presented in 1.1. In chapter 2, I examine formulations of Skolem's Paradox which are somewhat more sophisticated (relying, e.g., on the downward Löwenheim-Skolem theorem and/or the Mostowski collapsing lemma). Nothing in the present section (or chapter) is supposed to "solve" these more complicated paradoxes. In the long run, I argue that the solution given here *does* carry over—at least in its main outlines—to these other paradoxes, but I cannot give this argument until the alternate formulations of Skolem's Paradox are on the table. I will return to this topic near the end of chapter 2.

Third, it would be nice to have a solution to Skolem's Paradox which provides a little more detail about *which* of the semantic differences between $\Omega_E(x)$ and $\Omega_M(x)$ really leads to the failure of ($\dagger'$). That is, we might like to say that *this particular* instance of "$\in$" causes $\Omega_M(\hat{m})$ to be true and $\Omega_E(\hat{m})$ to be false or to identify *one particular* instance of "$\exists\, x$" whose differing interpretations "explain" the failure of ($\dagger'$). Unfortunately, I do not think that any such solution is available in general. Without a great deal more information about the particularities of the model $M$, there is no way to tell which of the *many* differences between $\Omega_E(x)$ and $\Omega_M(x)$ "explains" the failure of ($\dagger'$). To make this problem clear—and to examine some of the different ways ($\dagger'$) can fail—we need to examine several different versions of $M$. This examination will also take place in chapter 2.

Finally, the solution to Skolem's Paradox that I have given here is aimed primarily at the reader who is merely *puzzled* by the paradox—i.e., at the reader who finds Skolem's Paradox perplexing but who can't quite put his finger on where it goes wrong. For such readers, my isolation of $(\dagger')$ and my explanation of where $(\dagger')$ goes astray should help to demystify the paradox. On the other hand, readers who have a more theoretical commitment to Skolem's Paradox—e.g., readers with philosophical reasons for identifying the semantics of $\Omega_E(x)$ with those of $\Omega_M(x)$—will clearly find my solution less than satisfactory. I will deal with these more sophisticated—and more philosophically interesting—interpretations of Skolem's Paradox in chapter 3.

# Chapter 2

# Some Complicated Paradoxes

How could someone challenge the analysis of Skolem's Paradox presented in the last chapter? One way would be to use some more heavy-duty mathematics in the initial formulation of Skolem's Paradox. The hope, here, is that new mathematics will allow us to gain more control over the the fine structure of $M$, and that this additional control will allow us to make the semantics of $\Omega_M(x)$ closer to those of $\Omega_E(x)$. In the long run, the goal is to make these semantics *so* close that principles like ($\dagger'$) wind up being true.

Before discussing particular implementations of this strategy, I want to make a few comments concerning the strategy's general features and chances for success. First, because this is (essentially) a strategy for "patching up" argument (A), the arguments it produces tend to look a lot like argument (A). In particular, all of these arguments can be fit into the following, generic pattern:

> Using theorems $T_1, \ldots, T_j$, we obtain a model $M$ such that: 1.) $M$ is countable, 2.) $M \models ZFC$, and 3.) $M$ possesses "nice-making" properties $P_1, \ldots, P_k$. Now, because $M \models$ ZFC, there must be an $\hat{m} \in M$ such that $M \models \Omega[\hat{m}]$. Therefore:
>
> 1. $M$ is a countable model of ZFC which possesses $P_0, \ldots, P_k$.
> 2. $\Omega_E(x)$ says that "x is uncountable."
> 3. $M \models \Omega[\hat{m}]$.
>
> $\therefore$ 4. $\{x \mid x \in_M \hat{m}\}$ is uncountable.
>
> 5. If $M$ is countable and $m \in M$, then $\{x \mid x \in_M m\}$ is also countable.
>
> $\therefore$ 6. $\{x \mid x \in_M \hat{m}\}$ is countable.

Clearly, arguments fitting this pattern have a lot in common with argument (A): in particular, they still focus on countable models of ZFC, and they still emphasize the fact that such models' elements satisfy formulas like $\Omega(x)$. At the same time, their use of $T_1, \ldots, T_j$ and $P_1, \ldots, P_k$ provides additional information which makes (or which *purports* to make) them more plausible than the original argument (A).

Second, arguments following the pattern above can be analysed in much the same way as (A) itself was analysed. Let (X) be one such argument. As in 1.2.2, we can isolate a principle ($\dagger'_X$) which is both necessary and sufficient for (X) to be sound. This principle has the same outward form as ($\dagger'$) itself—i.e.

$$\forall m \in M \, [\Omega_M(m) \implies \Omega_E(\{x \mid x \in_M m\})]. \qquad (\dagger'_X)$$

However, whereas ($\dagger'$) involved the relatively nondescript model $M$ which was chosen at the beginning of this 1.1, $(\dagger'_X)$ involves the new $M$ which was chosen using $T_1, \ldots, T_j$ and which, in consequence, possess the nice-making properties $P_1, \ldots, P_k$.[1]

Once we have isolated $(\dagger'_X)$, we can continue to follow our earlier analysis and give two, fairly simple, arguments against this principle. On the one hand, we can note that the very $\hat{m}$ which was employed in formulating (X) provides a counterexample to $(\dagger'_X)$ (i.e., since $\Omega_M(\hat{m})$ is true while $\Omega_E(\{x \mid x \in_M \hat{m}\})$ is false). On the other hand, we can note that, even if we ignore the *details* of $(\dagger'_X)$, the fact that this principle makes (X) sound entails that there must be *something* wrong with it. These two arguments, along with the fact that $(\dagger'_X)$ is necessary for (X) to be sound, allow us to conclude that (X) fails.

Because this all works at a generic level—i.e., without knowing anything about the particular $T_1, \ldots, T_j$ and $P_1, \ldots, P_k$ involved in (X)—I will not run through this analysis for every argument in this chapter. Instead, I will assume that the basic *formulation* of arguments like (X) can be taken for granted and that the basic *refutation* of such arguments—i.e., that sketched in the last two paragraphs—needs no further elaboration. In place of such elaboration, then, I will focus on two more philosophically interesting aspects of arguments like (X): 1.) the exact role that theorems $T_1, \ldots, T_j$ and properties $P_1, \ldots, P_k$ play in making the relevant version of $(\dagger'_X)$ *seem* plausible and 2.) the ways in which $\Omega_M(x)$ and $\Omega_E(x)$ continue to differ even in the presence of $P_1, \ldots, P_k$.

Before I begin, a final comment is in order. Up until now, I have assumed very little set theory over-and-above the axioms of ZFC.[2] For the remainder of the chapter, I want to assume a little more. In particular, I will use ZFC + "there exists an inaccessible cardinal, $\kappa$" as my background set theory throughout sections 2.1 and 2.2. Basic information about inaccessible cardinals can be found in [26] and [30] or in sections 1.1 and 4.3.1 of this dissertation.

## 2.1 Transitivity

In section 1.1, I introduced the notion of a *transitive* model for the language of set theory. In this section, I look in more detail at the role these models play (or *can* play) in the formulation of Skolem's Paradox. I begin by recalling the definition of transitivity. We say that a model $M$ is transitive when the model's domain is a transitive set (i.e., when all of the members of $M$ are sets and when every member of a member of $M$ is also a member of $M$) and the "membership" relation on $M$ is just the real membership relation restricted to $M$'s domain (i.e., when $i_M(\in) = \{\langle m_1, m_2 \rangle \in M \times M \mid m_1 \in m_2\}$).

Given this definition, how do we get a transitive model of ZFC? If we follow the argument sketched on page 8 (cf. fn. 8), it seems that there are three things we should use. First, we use the fact that $\kappa$ is an inaccessible cardinal to get that $\langle V_\kappa, \in \rangle \models$ ZFC. Second, we use the so-called "downward Löwenheim-

[1] In theory, we could eliminate any possible confusion here by employing a whole string of primes: $M$, $M'$, $M''$, etc. However, since the particular model under discussion will always be clear from context, I will avoid this kind of notational complexity.

[2] I have, of course, assumed that ZFC is consistent, as this assumption is necessary to obtain models of ZFC! This, however, is a fairly *weak* extension of ZFC, and without it Skolem's Paradox becomes rather boring.

Skolem theorem" to get a countable, elementary submodel of $\langle V_\kappa, \in \rangle$ (call this model $N$). Finally, we use the Mostowski collapsing lemma (see p. 8, fn. 8) to get a transitive $M$ which is isomorphic to $N$.

At the end of this construction, there are two things to notice. First, $M$ is clearly a transitive model for the language of set theory. Second, the fact that $N \prec \langle V_\kappa, \in \rangle$ ensures that $N$ and $\langle V_\kappa, \in \rangle$ are elementarily equivalent (see p. 7, fn. 6), and the fact that $N$ and $M$ are isomorphic ensures that $N$ and $M$ are elementarily equivalent. Hence, $\langle V_\kappa, \in \rangle$ and $M$ are elementarily equivalent as well. In particular, $M \models$ ZFC.[3]

This, then, gives us a way of obtaining a countable, transitive model of ZFC. Let's call the version of Skolem's Paradox which results from employing this model "Argument (T)," and let's see what the transitivity of $M$ does for (T). As far as I can see, there are two things which transitivity does for argument (T)—i.e., two advantages which (T) has over the original, unadorned argument (A).

First, transitivity eliminates the interpretive problem which plagued us in section 1.1. There, we were forced to conclude that proponents of Skolem's Paradox use the phrase "x is uncountable" to mean that $\{y \mid y \in_M x\}$ is uncountable. And, while this gives a good reading of Skolem's Paradox, it's also somewhat unnatural. The natural way to use "x is uncountable" is to mean that $\{y \mid y \in x\}$ is uncountable. This tension between the natural reading of "x is uncountable" and the reading which makes Skolem's Paradox work (or at least makes premise 5 in argument (A) work!) created somewhat of a certain strain in our overall interpretation of Skolem's Paradox.

For transitive $M$, however, this strain goes away. Since the membership relation on a transitive model *just is* the ordinary membership relation, there is no difference between the sets $\{y \mid y \in_M x\}$ and $\{y \mid y \in x\}$. As a result, we can go ahead and use the "natural" reading of "x is uncountable" while availing ourselves of all the benefits which accrue to the $\{y \mid y \in_M x\}$ reading. This gives us a first advantage of using transitive models to formulate Skolem's Paradox.

The second advantage of using transitive models comes from the elimination of (some of the) semantic differences between $\Omega_M(x)$ and $\Omega_E(x)$. At the most basic level, moving to transitive models eliminates extensional differences involving the interpretation of "$\in$." As we saw in section 1.3.3, for arbitrary $N$ there need be no connection—or even extensional overlap—between the interpretation of "$\in$" in $\Omega_N(x)$ and the interpretation of "$\in$" in $\Omega_E(x)$. For *transitive* $N$, however, these two interpretations coincide: for any $n_1, n_2 \in N$, $n_1 \in_N n_2 \iff n_1 \in n_2$. In the case of transitive models, therefore, one of the fundamental differences between the semantics of $\Omega_M(x)$ and those of $\Omega_E(x)$ simply disappears.

At a more sophisticated level, the transitivity of $M$ actually ensures that $M$ "gets it right" about a lot more than just the membership relation. Let me say that a relation $R$ is *absolute* for transitive models if

[3] From the standpoint of consistency strength, this construction involves quite a bit of overkill. The assumption that transitive models of ZFC exist is *far* weaker than the assumption that inaccessible cardinals exist. Hence, we could have obtained such a model without employing inaccessible cardinals. However, since we *are* assuming the existence of a inaccessible, the construction sketched above is a perspicuous way to get transitive models.

there is some formula $\Psi^R(x)$ such that for any transitive $N \models$ ZFC and any $\bar{n} \in N$:

$$R \text{ holds of } \bar{n} \iff \Psi^R_E(\bar{n}) \iff \Psi^R_N(\bar{n}) \iff N \models \Psi^R[\bar{n}].$$

The definition of transitive models ensures that the relation "is a member of" is absolute for transitive models. With a bit more work, we can also show that the following are absolute:

- $f$ is a function; $f$ is injective; $f$ is surjective; $f$ is bijective.
- $x = \text{Domain}(f)$; $x = \text{Range}(f)$.
- $x$ is finite; $x$ is infinite; $x$ is an ordinal; $x$ is a limit ordinal; $x = \omega$.

Hence, transitive models "know" quite a lot about the sets they contain. For a wide range of set-theoretic concepts, transitive models of ZFC can pin these concepts down in a way which is guaranteed to be accurate (at least, that is, for elements in the models' domains).

These, then, give two ways in which transitivity helps to make Skolem's Paradox more plausible. By making $M$ a transitive model, we clear up some awkwardness in our initial formulation of premise 5. We also ensure that $M$ really does capture a lot of set theoretic concepts—i.e., that for a wide and interesting range of formulas $\Psi$, $\Psi_M$ and $\Psi_E$ are extensionally equivalent. Our question, then, is why this result doesn't carry over to the particular case we are interested in—i.e., to the case of $\Omega_M(x)$ and $\Omega_E(x)$.

The answer to this question comes in four stages. At the most rudimentary level, the discussion up to this point entails that any difference between $\Omega_M(x)$ and $\Omega_E(x)$ must be located in the interpretation of "$\exists x$." After all, the only extensional differences between $\Omega_E(x)$ and our original $\Omega_M(x)$—i.e., the $\Omega_M(x)$ which used the $M$ chosen at the beginning of this 1.1—involved the interpretation of "$\in$" and "$\exists x$." Since our new $M$ is transitive, any differences involving "$\in$" have been eliminated. Therefore, the remaining differences must involve the interpretation of "$\exists x$."

Moving a little deeper, we find that the transitivity of $M$ allows us to isolate *just which instance* of "$\exists x$" leads to the crucial difference between $\Omega_M(x)$ and $\Omega_E(x)$. To see this, we need to get a little more machinery on the table. We can begin by reexamining the list of absolute concepts which appeared on page 33. In doing so, we find that this collection is rich enough to include the two-place relation "$f$ is a bijection between $x$ and $\omega$." In other words, there there exists a formula $\Psi(f, x)$ in the language of set theory such that for any transitive $N \models$ ZFC and any $n_1, n_2 \in N$,

$$n_1 \text{ is a bijection between } n_2 \text{ and } \omega \iff \Psi_E(n_1, n_2) \iff \Psi_N(n_1, n_2).^4$$

Further, our original formula $\Omega(x)$ is closely related to this $\Psi(f, x)$. Specifically,

$$\Omega(x) \equiv_{df} \neg\exists f\, \Psi(f, x).$$

[4] To get this $\Psi$, note that we already have formulas $\Psi_1(f)$, $\Psi_2(x, f)$, $\Psi_3(y, f)$, and $\Psi_4(y)$, which capture, respectively, the concepts "$f$ is a bijection," "$x = \text{Domain}(f)$," $y = \text{Range}(f)$, and "$y = \omega$." Therefore, the formula:

$$\Psi(f, x) \equiv_{df} \Psi_1(f) \wedge \Psi_2(x, f) \wedge \exists y\, (\Psi_3(y, f) \wedge \Psi_4(y))$$

must capture the concept "$f$ is a bijection between $x$ and $\omega$."

This, then, provides us with all the machinery we need to determine where the two sentences $\Omega_E(x)$ and $\Omega_M(x)$ really differ.

To begin, note that the sentences $\Omega_E(x)$ and $\Omega_M(x)$ clearly interpret the symbol "$\neg$" in the same way. Next, note that the absoluteness of $\Psi(f, x)$ ensures that, for any particular $f, x \in N$, the sentences $\Psi_E(f, x)$ and $\Psi_M(f, x)$ are also (extensionally) equivalent. Combining these two facts, we see that any difference between the semantics of $\Omega_E(x)$ and $\Omega_M(x)$ must involve the interpretation of the existential quantifier highlighted in "$\neg\exists f\, \Psi(f, x)$"—i.e., the leftmost existential quantifier in $\Omega(x)$. This quantifier, therefore, is the one which "explains" the ultimate failure of $(\dagger'_M)$.

Let's look a little closer at this "explanation." Since we already know that the case $x = \hat{m}$ is an instance of the failure of $(\dagger'_M)$—i.e., that the conditional $\Omega_M(\hat{m}) \implies \Omega_E(\hat{m})$ is both false and an instantiation of $(\dagger'_M)$, we can focus our attention on this case. Given what we know about $\Psi(x)$, the following two facts should be clear:

1. For any set $f$, $\Psi_E(f, \hat{m})$ is true if and only if $f$ is a bijection between $\hat{m}$ and $\omega$.
2. For any $f \in M$, $\Psi_E(f, \hat{m})$ is true if and only if $\Psi_M(f, \hat{m})$ is true.

Further, the fact that $M$ is countable entails that $\{m \in M \mid m \in \hat{m}\}$ is countable—i.e., that there really is a bijection $\hat{f} : \hat{m} \to \omega$.

These facts highlight the relevance of the initial quantifier in $\Omega(x)$ to the failure of $(\dagger'_M)$. Since there really is a bijection $f : \hat{m} \to \omega$, and since the initial quantifier in $\Omega_E(\hat{m})$ ranges over the universe of sets—i.e., over a domain large enough to contain $\hat{f}$, this quantifier "recognizes" the fact that $\Psi_E(\hat{f}, \hat{m})$ is true. As a result, $\Omega_E(\hat{m})$ comes out false. In contrast, neither $\hat{f}$ nor any other bijection between $\hat{m}$ and $\omega$ actually lives in the domain of $M$. Hence, the initial quantifier in $\Omega_M(\hat{m})$ doesn't find any $f$ for which $\Psi_M(f, \hat{m})$ is true, and so $\Omega_M(\hat{m})$ winds up being true.

This point can be put in somewhat more ordinary English. Basically, all my talk about "absoluteness" is a means of ensuring that the following are reasonable paraphrases of $\Omega_E(\hat{m})$ and $\Omega_M(\hat{m})$ respectively:[5]

1. There is no $f$ in the set-theoretic universe such that $f : \hat{m} \to \omega$ is a bijection.
2. There is no $f \in M$ such that $f : \hat{m} \to \omega$ is a bijection.

Given these paraphrases, the explanation for $(\dagger'_M)$'s failure cannot be hard to find. Since $\hat{m}$ really is countable, there is a bijection between $\hat{m}$ and $\omega$; hence 1 is false. This, however, does not entail that 2 is false. As long as none of the $f$'s which falsify 1 happen to live in the domain of $M$, 2 can (and does) continue to be true. This, then, is the "plain English" explanation for $(\dagger'_M)$'s failure: even though there exist bijections from $\hat{m}$ to $\omega$—and so, bijections which witness the failure of $\Omega_E(\hat{m})$—none of these bijections lives inside the domain of $M$. So, $\Omega_M(\hat{m})$ still winds up being true.

[5] Here, the absoluteness of $\Psi$ serves to ensure that the phrase "$f : \hat{m} \to \omega$ is a bijection" can be used fairly (and literally) in cashing out *both* $\Omega_E(\hat{m})$ and $\Omega_M(\hat{m})$.

There are two things which are worth mentioning about this explanation for the failure of ($\dagger'_M$). First, our ability to pin down the particular quantifier which accounts for this failure *depends* on the fact that $M$ is a transitive model. It is *because* $M$ is transitive that we know that $\Psi_E(f,x)$ and $\Psi_M(f,x)$ are equivalent, and it is because we know this equivalence that we can be confident that the initial "$\exists x$" in $\Omega(x)$ is the place where $\Omega_E(x)$ and $\Omega_M(x)$ ultimately disagree. If $M$ were not transitive, then any of the (many) instances of "$\exists x$" and "$\in$" which occur in $\Psi(f,x)$ could just as well "explain" the failure of ($\dagger'_M$)—i.e., could explain it just as well as our particular "$\exists x$" does.

Second, there is something which may still seem a bit paradoxical about this whole analysis. As we saw in our initial discussion of $\Omega_E(x)$, the concept of uncountability can be *defined* using only the notion of membership and terms from ordinary classical logic. How, then, can any semantics which gets the one non-logical piece of machinery in this definition "right"—i.e., a semantics which interprets "$\in$" in accordance with ordinary English usage—still diverge from ordinary English in its understanding of uncountability? Shouldn't the fact that we use *purely logical* machinery to define "uncountable" in terms of "$\in$" ensure that any two semantics which agree on "$\in$" also agree on "uncountable"?

To see what's wrong with this line of thought, we need to notice two things. First, even in ordinary English, phrases which use quantifiers are often allowed to vary according to local context. Consider, for instance, the ordinary English sentence

$$\text{Owen is the tallest person.} \tag{O}$$

Depending on the context in which this sentence is uttered, it can mean different things. In some contexts, it means that Owen is the tallest person *in the conversation,* in others, that Owen is the tallest person *in the room,* and in still others, that Owen is the tallest person *in the world.* The differences between these interpretations of (O) don't involve different understandings of who Owen is; nor do they involve different conceptions of tallness. Instead, they involve different assumptions about the relevant "domain of discourse"—i.e., about the class of people to whom Owen is being compared.

To connect this point more closely to our discussion of $\Omega_E(x)$ and $\Omega_M(x)$, let's modify the example slightly and consider the following definition of "Tallest" (phrased in terms of the concept "Taller"):[6]

$$\text{Tallest}(x) \equiv_{df} \neg\exists y \ \text{Taller}(y,x). \tag{O$_1$}$$

As with the case of (O) above, the significance of this definition varies depending on the (contextually determined) range of the initial existential quantifier. In some contexts, the definition suffices to make me the "tallest" person (say, if I give the definition while on a walk with my wife and daughter). In other contexts, it does not (say, when I give the definition while lecturing to an assorted group of Division I basketball players).

The key point, then, is this: the mere fact that O$_1$ is formulated using "logical vocabulary" does not ensure that it has a constant significance across all contexts (even across all contexts where the non-logical

[6] The reader will note that this definition of Tallest has *exactly* the same form as our canonical definition of "uncountable."

vocabulary in $\mathrm{O}_1$ retains a fixed significance). Even in ordinary English, quantifiers function in different ways depending on the (contextually determined) collections over which they are allowed to range. As a result, definitions which use quantifiers will pick out different objects (or classes) depending on the range these quantifiers have (even in cases where the rest of the symbols in the definitions have fixed meanings). Given this, we should not be surprised to find that this phenomenon shows up in the case where "uncountability" happens to be the definandum in question.

This gives a first reason for thinking that there is nothing to the idea that, insofar as all the symbols in $\Omega(x)$ are either "$\in$" or logical constants, fixing the interpretation of "$\in$" *must* fix the interpretation of $\Omega(x)$. To amplify this point, we should notice two facts about modern model theory. First, the definition of satisfaction in modern model theory is clearly *designed* to allow variation in the range of quantifiers. For models $N_1$ and $N_2$, and for any formula $\Phi(x)$, the definition of satisfaction says that:

$$N_1 \models \exists x \Phi(x) \Longleftrightarrow \text{there exists an } n \in N_1 \text{ such that } N_1 \models \Phi[n].$$
$$N_2 \models \exists x \Phi(x) \Longleftrightarrow \text{there exists an } n \in N_2 \text{ such that } N_2 \models \Phi[n].$$

Given this, there is no reason to *expect* "$N_1 \models \exists x \Phi(x)$" and "$N_1 \models \exists x \Phi(x)$" to have the same truth-value. Even in cases where $N_1$ and $N_2$ agree about all elements in their common domain (i.e., even where $N_1 \models \Phi(n) \Leftrightarrow N_2 \models \Phi(n)$ for every $n \in N_1 \cap N_2$), the fact that the recursion clauses for $\exists$ evaluate different *classes* of $n$'s makes it entirely reasonable to suspect that "$N_1 \models \exists x \Phi(x)$" and "$N_1 \models \exists x \Phi(x)$" have different truth-values.

Second, this difference in the way quantifiers are evaluated *often* leads to the kind of phenomena we are investigating here. Consider, for instance, the following two models for the language containing only a single, unary function $s$:

- Domain$(N_1) = \mathbb{N}$; $s_1 = \{\langle n, n+1 \mid n \in N_1 \rangle\}$.
- Domain$(N_2) = \mathbb{N} \setminus \{0\}$; $s_1 = \{\langle n, n+1 \mid n \in N_2 \rangle\}$.

Next, consider the obvious way to define "Zero" in these models:

$$\text{Zero}(n) \equiv_{df} \neg \exists x \, [s(x) = n].$$

Despite the fact that these two models contain almost exactly the same members (differing only in that $N_1$ contains 0 while $N_2$ does not), and despite the fact that the models interpret "$s$" in almost exactly the same way (agreeing on the value of $s(n)$ for all $n \in N_2$), this definition clearly functions differently in the two models. In particular, $N_2 \models \text{Zero}[1]$ while $N_1 \models \neg\, \text{Zero}[1]$. The difference here is *purely* a result of the different ranges over which the initial quantifier in our definition "Zero" is allowed to range. In one case, the quantifier ranges over a domain which includes 0, so it makes Zero[1] false. In the other, the quantifier's domain is sufficiently limited that it never finds an $n$ such that $s(n) = 1$; hence, Zero[1] comes out true.

To emphasize the utter ordinariness of this situation, note that everything I just said about the definition of "Zero" applies equally to the definitions of "One," "Two," etc.[7] Further, if we continue to construct new models by eliminating initial elements from $\mathbb{N}$ (i.e., letting Domain($N_3$) = $\mathbb{N} \setminus \{0, 1\}$, Domain($N_4$) = $\mathbb{N} \setminus \{0, 1, 2\}$, etc.), we obtain a whole chain of models $N_1 \subset N_2 \subset N_3 \ldots$ such that 1.) for any $i < j$, $N_i \setminus N_j$ is finite, 2.) for any $n \in N_i \cap N_j$, $s_i(n) = s_j(n)$ and 3.) *no* two $N_i$'s agree on the definition of Zero or One, or Two, etc. That is, if we let $i < j$, let $n \in N_i \cap N_j$ and let $\Phi$ represent one of the formulas, Zero, One, Two, etc., then

$$N_j \models \Phi[n] \Longrightarrow N_i \not\models \Phi[n].$$

Hence, even in the case of quite simple models (and quite simple definitions), the type of phenomenon which appeared in our analysis of $\Omega_M(x)$—i.e., the phenomenon of a single quantifier "explaining" differences in the way a definition really functions—appears to be quite commonplace.

This, then, gives a fairly full analysis of the role transitive models play (or can play) in the formulation of Skolem's Paradox. As we saw earlier, the insistence that $M$ be transitive allows us to make the semantics of $\Omega_M(x)$ and $\Omega_E(x)$ closer in many interesting and consequential ways. In the long run, however, differences between the way $\Omega_M(x)$ and $\Omega_E(x)$ interpret their initial existential quantifier are substantial enough to lead to a failure of ($\dagger'_M$) (and, hence, to a failure of argument $T$). Further, we have seen that there is nothing particularly surprising (still less *paradoxical*!) about the fact that a single quantifier can play this role. In model theory, this kind of situation occurs all the time; and, even in ordinary English, variation in the domain of a quantifier often leads to differences in truth value for sentences in which that quantifier occurs.

Before closing, I want to re-emphasize the fact that it is the transitivity of $M$ which lets us give this particular analysis of the failure of ($\dagger'_M$)—i.e., which lets us isolate the initial quantifier in $\Omega(x)$ as *the* symbol which "explains" the difference in truth-value between $\Omega_M(x)$ and $\Omega_E(x)$. In the next section, I examine several cases in which this analysis does not work. In doing so, I illustrate the fact that ($\dagger'_M$) can fail in many *different* ways and that, as a result, there can be no "uniform" solution to Skolem's Paradox.

## 2.2 Elementarity

Given what we have seen in the last section, it might be tempting to think that *all* versions of Skolem's Paradox can be solved in the same manner as argument (T)—i.e., by examining differences in the way $\Omega_E(x)$ and $\Omega_M(x)$ interpret the initial existential quantifier in $\Omega(x)$. This thought clearly has two things going for it. First, there are *some* models for which this analysis really does provide the best solution to Skolem's Paradox—e.g., the transitive models discussed in the last section. Second, so long as $M$ is countable, $\Omega_E(x)$ and $\Omega_M(x)$ *do* interpret their initial existential quantifiers differently (since $\Omega_E(x)$ reads it as "there is a set" and $\Omega_M(x)$ reads it as "there is a set *in M*").

Despite these points, I don't think a focus on initial quantifiers provides an adequate understanding of

[7] E.g., when we define One($n$) $\equiv_{df}$ $\exists x\, s(x) = n \wedge \neg\exists x\, s(s(x)) = n$ and Two($n$) $\equiv_{df}$ $\exists x\, s(s(x)) = n \wedge \neg\exists x\, s(s(s(x))) = n$, etc.

what's really going on in Skolem's Paradox. To see why, we can begin by looking at a model in which the "explanation" for the failure of ($\dagger'_M$), though clearly involving the interpretation of *quantifiers*, does not involve the interpretation of *initial quantifiers*. As in section 2.1, we begin by letting $\kappa$ be an inaccessible cardinal and letting $M$ be a countable elementary submodel of $V_\kappa$. There are two things to notice about this model: 1.) $M$ is clearly a countable model of ZFC and 2.) for any $\Psi(\bar{x})$ in the language of set theory and any $\bar{m} \in M$

$$M \models \Psi[\bar{m}] \iff V_\kappa \models \Psi[\bar{m}].$$

It is this second property of $M$ which will be of the most interest to us here. I begin, therefore, by saying a few things about the ways this property helps make ($\dagger'_M$) appear plausible.

First, note that $M$'s being an elementary submodel of a transitive model ensures that $M$ inherits all of the absoluteness properties which were discussed in the last section. So, let $R(\bar{x})$ be a property which is absolute for transitive models and let $\Psi^R(\bar{x})$ be the formula which witnesses this absoluteness. Then for any $\bar{m} \in M$,

$$R \text{ holds of } \bar{m} \iff \Psi^R_E(\bar{m}) \iff V_\kappa \models \Psi^R[\bar{m}] \iff M \models \Psi^R[\bar{m}] \iff \Psi^R_M(\bar{m}). \tag{K}$$

Here, the first two biconditionals come from the absoluteness of $R$, the third from the fact that $M \prec V_\kappa$, and the last from the definition of $\models$ (see p. 17).

Second, the fact that $\kappa$ is inaccessible ensures that, in addition to all the properties discussed in the last section, the property of being uncountable is also absolute for $V_\kappa$. That is, for any $n \in V_\kappa$,

$$n \text{ is uncountable } \iff \Omega_E(n) \iff V_\kappa \models \Omega[n].$$

When this fact is combined with the one discussed above, we obtain the key fact about $M$ and uncountability: for any $m \in M$,

$$m \text{ is uncountable } \iff \Omega_E(m) \iff \Omega_M(m).$$

As a result, $M$ really does "get it right" about uncountability. The fact that $M$ is an elementary submodel of $V_\kappa$, combined with the fact that "being uncountable" is absolute for $V_\kappa$ ensures that "being uncountable" is also absolute for $M$.

Why, then, doesn't this make ($\dagger'_M$) trivially true? There are two things to be said here. First, let $\hat{m}$ be a member of $M$ such that $M \models \Omega[\hat{m}]$. Then, in contrast to the transitive case, $\hat{m} \neq \{m \,|\, m \in_M \hat{m}\}$. In particular, although every $m' \in \hat{m} \cap M$ is also in $\{m \,|\, m \in_M \hat{m}\}$, there are quite a few (indeed, uncountably many!) $m' \in \hat{m} \setminus M$ (hence, in $\hat{m} \setminus \{m \,|\, m \in_M \hat{m}\}$). As a result, there is no reason to think that $\Omega_E(\{m \,|\, m \in_M \hat{m}\})$ and $\Omega_E(\hat{m})$ are equivalent, so there is no reason to think that ($\dagger'_M$) follows from the biconditionals in the last paragraph. This gives us a fairly simple answer to the question with which this paragraph began.

Moving somewhat deeper, we should note that the difference between $\Omega_M(\hat{m})$ and $\Omega_E(\{m \,|\, m \in_M \hat{m}\})$ cannot, in this case, be explained by appealing to differing interpretations of the initial existential quantifier in $\Omega(x)$.[8] To be sure, the quantifiers in $\Omega_E(x)$ *do* range over a wider domain than those in $\Omega_M(x)$, and some (uncountably many!) of the objects in this wider domain are genuine bijections between $\{m \mid m \in_M \hat{m}\}$ and $\omega$. However, this fact is not what *explains* the difference between $\Omega_M(\hat{m})$ and $\Omega_E(\{m \,|m \in_M \hat{m}\})$.

To see this, we need to reexamine the rational for thinking that the initial quantifier in $\Omega(x)$ *could* explain this difference in the first place. Presumably, the thought goes something like this. There is some particular function $f$ (or perhaps some class of functions $\mathcal{F}$) which has the following properties:

1. $f$ lives in the range of the initial quantifier in $\Omega_E(\{m \,|m \in_M \hat{m}\})$.

2. $f$ does not live in the range of the initial quantifier in $\Omega_M(\hat{m})$.

3. $\Psi_E(f, \{m \,|m \in_M \hat{m}\})$ is true.

4. If $f$ *were* a member of $M$, then $\Psi_M(f, \hat{m})$ *would be* true.[9]

Given these properties, we reason that *if* the initial quantifier in $\Omega_M(\hat{m})$ could only "know" about $f$, then $\Omega_M(\hat{m})$ and $\Omega_E(\{m \,|m \in_M \hat{m}\})$ would agree. Hence, we conclude that it's the difference in the way $\Omega_M(\hat{m})$ and $\Omega_E(\{m \,|m \in_M \hat{m}\})$ interpret their initial quantifiers which "explains" the failure of $(\dagger'_M)$.

In the case of $M$, however, it's hard to see what "$f$" is supposed to represent. On the one hand, we could look for some bijection $f : \hat{m} \to \omega$. Using the absoluteness properties of $M$, we could make a plausible case that this $f$ would have property 4.[10] Unfortunately, such an $f$ would not have property 3.[11] Still more unfortunately, it's easy to show that no such $f$ even exists! From what we have already seen , we know that

$$\hat{m} \text{ is uncountable } \iff \Omega_M(\hat{m}).$$

Hence, the very fact that $M \models \Omega[\hat{m}]$ ensures that there is no bijection $f : \hat{m} \to \omega$. This, together with the fact that such an $f$ wouldn't satisfy 3 anyway, shows that looking for *this kind* of bijection is a mistake.

On the other hand, we could shift our focus and look for a bijection $f : \{m \mid m \in_M \hat{m}\} \to \omega$. Such a bijection would clearly avoid the problems just discussed: it would be relatively easy to find, and it would satisfy properties 1–3. However, it is not clear why this bijection should satisfy property 4. After all, the very

[8] Note that the argument of the last paragraph only explained why the biconditionals from the penultimate paragraph don't *entail* $(\dagger'_M)$. It doesn't explain why $(\dagger'_M)$ is *wrong*. Explaining this is a different, and somewhat more delicate, task.

[9] Clearly, it's hard to figure out the significance of this necessarily-contrary-to-fact conditional. Nevertheless, I think that this conditional, or something very much like it, provides the best formulation of the idea that initial quantifiers *explain* the difference between $\Omega_E(\{m \,|m \in_M \hat{m}\})$ and $\Omega_M(\hat{m})$.

[10] Here, the absoluteness properties of $M$ ensure that, for any $m \in M$,

$$m \text{ is a bijection between } \hat{m} \text{ and } \omega \iff \Psi_M(m, \hat{m}).$$

Hence, and modulo the concerns mentioned in the last footnote, we could conclude that, *if* $f$ were a member of $M$, then $\Psi_M(f, \hat{m})$ would have to be true.

[11] Since the domain of $f$ is $\hat{m}$ rather than $\{m \mid m \in_M \hat{m}\}$, and since $\Psi_E(x, y)$ expresses the notion that $x$ is a function between $y$ and $\omega$, $\Psi_E(f, \{m \mid m \in_M \hat{m}\})$ cannot be true. Hence, 3 fails for $f$.

absoluteness properties which make ($\dagger'_M$) seem plausible, also entail that, for any $f : \{m \mid m \in_M \hat{m}\} \to \omega$,

$$\Psi_M(f, \hat{m}) \Longleftrightarrow \hat{m} = \{m \mid m \in_M \hat{m}\}.$$

Hence, the fact that $\hat{m} \neq \{m \mid m \in_M \hat{m}\}$ seems to entail that $\Psi_M(f, \hat{m})$ can't be true—i.e., that $\Psi_M(f, \hat{m})$ wouldn't be true *even if* $f$ were a member of $M$. If this is right, then looking for an $f : \{m \mid m \in_M \hat{m}\} \to \omega$ would be just as much of a mistake as looking for an $f : \hat{m} \to \omega$.

Note that this second example actually shows that *no* $f$ can play the role required by the argument sketched above. If $f$ is not a bijection between $\{m \mid m \in_M \hat{m}\}$ and $\omega$, then 3 will not be true. If $f$ *is* a bijection between $\{m \mid m \in_M \hat{m}\}$ and $\omega$, then the fact that $\hat{m} \neq \{m \mid m \in_M \hat{m}\}$ makes it highly implausible to think that $\Psi_M(f, \hat{m})$ "would be true" if only $f$ were in $M$. In no case, therefore, can we find a function $f$ which gives content to the idea that differences between $\Omega_M(\hat{m})$ and $\Omega_E(\{m \,|m \in_M \hat{m}\})$ are "explained" by appealing to differing interpretations of the initial quantifier in $\Omega(x)$.

Of course, the differences between $\Omega_M(\hat{m})$ and $\Omega_E(\{m \,|m \in_M \hat{m}\})$ *are* explained by differing interpretations which these sentences give to their quantifiers (since, after all, $\Omega_M(x)$ and $\Omega_E(x)$ agree on the interpretation of all other symbols in $\Omega(x)$—i.e., on "$\in$," "$\neg$," and "$\to$"). But, the quantifiers in question are those embedded in the formula $\Psi(f, x)$ and not those which stand at the beginning of $\Omega_M(x)$ and $\Omega_E(x)$.[12] Hence, although it might be tempting to give a uniform explanation for the failure of ($\dagger'_M$) in terms of the differing interpretations $\Omega_M(x)$ and $\Omega_E(x)$ give to their initial quantifiers, this uniform explanation cannot, in the long run, be sustained.

## 2.3 Membership

In this section, I want to continue exploring the main theme of the last section—the theme that initial quantifiers can't explain all instances of Skolem's Paradox—by examining a situation in which quantifiers *in general* don't explain what's really going on in Skolem's Paradox. To set this situation up, we need to construct (yet) another model. We begin by letting $N$ be a countable, transitive model of ZFC, letting $\hat{m}$ be an element of $N$ such that $N \models \Omega[\hat{m}]$, and letting $\hat{n}$ be an arbitrary element of $N$ such that $\text{Rank}(\hat{n}) > \text{Rank}(\hat{m}) + \omega$.[13] Next, we use the fact that $\hat{m}$ is countable to get a bijection $f : \hat{m} \to \omega$, and we

[12] Note that it is these internal quantifiers which differentiate the formulas $\Psi_E(f, \{m \mid m \in_M \hat{m}\})$ and $\Psi_M(f, \hat{m})$ that have been so crucial through the last few paragraphs.

[13] A few remarks on this choice of $\hat{n}$ are probably in order. Basically, I have chosen $\hat{n}$ so as to ensure that $\hat{n}$ does not live in the same "part" of $N$ as $\hat{m}$. In particular, $\hat{n}$ is not a member of either $\hat{m}$ or $\omega$. What's more, $\hat{n}$ is not equal to any ordered pair of the form $\langle n_1, n_2 \rangle$, where $n_1 \in \hat{m}$ and $n_2 \in \omega$, nor is $\hat{n}$ equal to any collection of such ordered pairs. As a result, we can manipulate $\hat{n}$ in various ways without modifying the parts of $N$ which directly involve $\hat{m}$, $\omega$, and $\hat{n} \times \omega$. The significance of this choice of $\hat{n}$ will become clear as my argument progresses.

define a function $\sigma : N \to (N \setminus \{\hat{n}\}) \cup \{f\}$ such that:

$$\sigma(n) = \begin{cases} n & \text{if } n \neq \hat{n} \\ f & \text{if } n = \hat{n} \end{cases}$$

Finally, we use $\sigma$ to define a new model, $M$, such that Domain$(M) = (N \setminus \{\hat{n}\}) \cup \{f\}$ and $\sigma$ is an isomorphism between $N$ and $M$ (see fn. 6 on p. 7 for the details of this construction).

There are four things which we should observe about the $M$ which results from this construction. First, because $M$ is isomorphic to $N$, $M$ is still a countable model for $ZFC$. Second, and again because $\sigma : N \to M$ is an isomorphism, $M \models \Omega[\hat{m}]$. Third, the fact that $M$ is countable ensures that there is *some* bijection $f' : \hat{m} \to \omega$ such that $f' \notin M$ (since there are, after all, $2^{\aleph_0}$ different bijections between $\hat{m}$ and $\omega$, not all of which can live in the countable model $M$). Finally, the fact that $f \in M$ means that $M$ contains *at least one* function which witnesses the fact that $\hat{m}$ is countable. Even though $M$ may not "recognize" this function in the right sort of way—e.g., as indicated by the fact that $M \models \Omega[\hat{m}]$—$M$ *does contain* this function.

What are we to make of this example? On my reading, the example provides an instance of Skolem's Paradox in which the failure of $(\dagger'_M)$ is due, not to different interpretations of the initial existential quantifier in $\Omega(x)$, but to different interpretations of the "$\in$" symbol in the embedded $\Psi(f, x)$. To see this, we should begin by noticing that, because $M \models \Omega[\hat{m}]$, $\Psi_M(f, \hat{m})$ must be false. In contrast, the fact that $f$ really is a bijection between $\hat{m}$ and $\omega$ entails that $\Psi_E(f, \hat{m})$ is true. These facts, together with the fact that both $\Omega_M(x)$ and $\Omega_E(x)$ "know" about $f$—i.e., that $f$ is within the range of the quantifiers in both $\Omega_M(x)$ and $\Omega_E(x)$—entail that any difference between $\Omega_M(x)$ and $\Omega_E(x)$ must be located in $\Psi(f, \hat{m})$ (and not, say, in the way these sentences interpret their initial existential quantifiers).

Let us, then, look more closely at $\Psi(f, \hat{m})$. Abbreviating wildly, we can write:

$$\begin{aligned} \Psi(f, x) \quad \equiv_{df} \quad & \forall x \in f\, [x \in \hat{m} \times \omega] \\ & \wedge \forall x \in \hat{m}\ \exists! y \in \omega\, [\langle x, y \rangle \in f] \\ & \wedge \forall y \in \omega\ \exists! x \in \hat{m}\, [\langle x, y \rangle \in f]. \end{aligned}$$

When we examine this formula closely, we observe that many of its subformulas receive equivalent interpretations under the semantics of $\Psi_E(f, \hat{m})$ and those of $\Psi_M(f, \hat{m})$; to use our earlier jargon, these subformulas are *absolute* between $V$ and $M$. In particular, we should observe that for any set $s$:[14]

1. $s \in \hat{m} \Longleftrightarrow M \models s \in \hat{m}$.
2. $s \in \omega \Longleftrightarrow M \models s \in \omega$.
3. $s \in \hat{m} \times \omega \Longleftrightarrow M \models s \in \hat{m} \times \omega$.
4. If $s' \in \hat{m}$, and $s'' \in \omega$, then $s = \langle s_1, s_2 \rangle \Longleftrightarrow M \models s = \langle s_1, s_2 \rangle$.

[14] Note, here, that it's our original choice of $\hat{n}$ which ensures that these four claims are true. By choosing $\hat{n}$ to be in a different part of $N$ than $\hat{m}$ and $\omega$, we ensure that replacing $\hat{n}$ with $f$ will not effect these properties.

Finally, we should observe that every $s \in \hat{m} \cup \omega \cup \hat{m} \times \omega \cup \hat{n}$ is within the range of the quantifiers of both $\Psi_E(f, \hat{m})$ and $\Psi_M(f, \hat{m})$.

Keeping these observations in mind, we can isolate three kinds of differences between $\Psi_E(f, \hat{m})$ and $\Psi_M(f, \hat{m})$. First, there are differences that occur *within* subformulas that are, themselves, absolute between $V$ and $M$—e.g., formal differences in the interpretation of quantifiers within expressions like "$x \in \hat{m} \times \omega$." Second, there are differences in the interpretation of quantifiers where 1.) these quantifiers are explicitly bounded as they occur in $\Psi(f, \hat{m})$ and 2.) these quantifiers have ranges, both as they occur in $\Psi_E(f, \hat{m})$ and as they occur in $\Psi_M(f, \hat{m})$, which include every element in *either* of the relevant bounding sets. So, for instance, the initial quantifier in "$\forall x \in f\ [x \in \hat{m} \times \omega]$" is bounded by the expression "$\in f$," and the ranges of the quantifiers in both $\Psi_E(f, \hat{m})$ and $\Psi_M(f, \hat{m})$ include $f \cup \{y \mid y \in_M f\}$. Similarly, the initial quantifiers in "$\forall x \in \hat{m}\ \exists! y \in \omega\ [\langle x, y \rangle \in f]$" are bounded by "$\in \hat{m}$" and "$\in \omega$," and the ranges of the quantifiers in both $\Psi_E(f, \hat{m})$ and $\Psi_M(f, \hat{m})$ include $\hat{m} \cup \{y \mid y \in_M \hat{m}\}$ and $\omega \cup \{y \mid y \in_M \omega\}$.

Clearly, these first two kinds of difference cannot explain the difference in truth-value between $\Psi_E(f, \hat{m})$ and $\Psi_M(f, \hat{m})$. Because differences of the first kind are isolated within subformulas whose truth-values are constant between $\Psi_E(f, \hat{m})$ and $\Psi_M(f, \hat{m})$, these differences cannot be where $\Psi_E(f, \hat{m})$ and $\Psi_M(f, \hat{m})$ really diverge. Similarly for differences of the second kind. Since none of the sets which live within the range of $\Psi_E(f, \hat{m})$'s quantifiers but not within the range of $\Psi_M(f, \hat{m})$'s quantifiers are relevant to the truth-values of formulas like $\forall x \in f\ [x \in \hat{m} \times \omega]$ or $\forall x \in \hat{m}\ \exists! y \in \omega\ [\langle x, y \rangle \in f]$, the differences between these quantifier's ranges cannot explain the difference in truth-values between $\Psi_E(f, \hat{m})$ and $\Psi_M(f, \hat{m})$.

This, then, brings us to the third difference between $\Psi_E(f, \hat{m})$ and $\Psi_M(f, \hat{m})$: their differing interpretations of the expression "$x \in f$." As this is the only difference which remains between $\Psi_E(f, \hat{m})$ and $\Psi_M(f, \hat{m})$, it must explain the difference in truth-values between these two sentences. Further, this explanation is relatively intuitive: because $M$ doesn't know what the *real* members of $f$ are, and because the notion "being a function between $x$ and $y$" gets defined *in terms of* a set's members, $M$ doesn't know that $f$ is a function between $\hat{m}$ and $\omega$. Hence, it's not surprising that $M$ fails to satisfy the formula which "expresses" this notion: i.e., that $M \not\models \Psi(f, \hat{m})$. This is a simple result of the discrepancy between $M$'s understanding of the membership relation on $f$ and the *real* membership relation on $f$.

We have, therefore, an analysis of the failure of $(\dagger'_M)$ which is quite different from our analyses of the earlier failures of $(\dagger')$. In each of the other cases, the explanation for the failure of $(\dagger')$ involved the different interpretations which $\Omega_E(\hat{m})$ and $\Omega_M(\hat{m})$ give to their quantifiers. In this case, the explanation involves the different interpretations which $\Omega_E(\hat{m})$ and $\Omega_M(\hat{m})$ give to the symbol "$\in$." Because this explanation *is* so different from our earlier explanations—and, I think, from the "standard line" on Skolem's Paradox—it is worth stopping to address two concerns which the explanation might raise.

First concern: why can't our explanation of the difference between $\Omega_E(\hat{m})$ and $\Omega_M(\hat{m})$ still rely on the ways these sentences interpret their initial existential quantifiers? After all, there are $2^{\aleph_0}$ bijections $g : \hat{m} \to \omega$ which do not live in the domain of $M$ (since the set of such bijections has cardinality $2^{\aleph_0}$ while $M$

is only countable). Why couldn't one of these other bijections "explain" the fact that $\Omega_E(\hat{m})$ is true while $\Omega_M(\hat{m})$ is false? To address this concern, we need to look at two things.

First, we need to recall how a function which doesn't live in $M$ *could* "explain" the failure of $(\dagger'_M)$. As we saw on page 39, there are four things that need to happen for such an explanation to work (or, at least, to be moderately plausible). The fourth of these is the (somewhat difficult to understand) conditional:

4. If $f$ were a member of $M$, then $\Psi_M(f, \hat{m})$ would be true.

In the case we are considering, this amounts to the claim that, *if* $M$ were to know about one of the above-mentioned bijections $g : \hat{m} \to \omega$, *then* $M$ would recognize this bijection *as a bijection*. As a result, $M$ would satisfy some formula of the form $\Psi[g, \hat{m}]$ and, in consequence, would fail to satisfy $\Omega[\hat{m}]$.

But why should we accept this? $M$ *already does* contain a bijection $f : \hat{m} \to \omega$, and $M$ *doesn't* recognize this bijection as a bijection (or, at least, not as the right kind of bijection to make $\Psi_M(f, \hat{m})$ true). Why should $M$ do any better when it comes to *other* bijections? In the case of transitive $M$—the case discussed in 2.1—$M$ really did get all bijections "right." Hence, it was at least plausible to think that $M$ would continue to get new bijections right, *if* $M$ could only know about such bijections.[15] But, once $M$ misunderstands *some* bijections, there's no reason to think that it will do better with respect to *other* bijections. Hence, there is no intuitive reason to think that the concern now at issue is really well-motivated.

Second, we should note that there are cases in which phenomena *like* Skolem's Paradox can *only* be explained in terms of the different interpretations which models give to the "$\in$" symbol. So, for instance, let $N$ be a countable, transitive model of ZFC and let $N' = N[G]$ be a generic extension of $N$ such that $\omega_1^N$ has been "collapsed" so as to have cardinality $\aleph_0$.[16] Next, let $X = \{n \in N \mid N \models \text{"Rank}(n) < \omega_\omega\text{"}\}$ and let $\sigma : N \to N'$ be a bijection such that $\sigma \restriction X = \text{Id}$. Finally, using the trick from footnote 6, let $M$ be a new model such that the domain of $M$ is the same as that of $N'$ and such that $\sigma$ becomes an isomorphism between $N$ and $M$.

At this point, we are in a position to formulate a puzzle very much like Skolem's Paradox except that it holds *between* $M$ and $N'$. To begin, note that the fact that $\sigma \restriction X = \text{Id}$ ensures that $M$ and $N'$ agree about the membership relation on $\omega_1^N$. That is,

$$\hat{m} = \{x \mid x \in_{N'} \omega_1^N\} = \{x \mid x \in_M \omega_1^N\}.^{17}$$

However, $M$ and $N'$ do not agree about the *countability* of $\omega_1^N$. By construction, $N' \models \neg\Omega[\omega_1^N]$. In contrast, the fact that $\sigma : N \to M$ is an isomorphism and that $\sigma(\omega_1^N) = \omega_1^N$ ensures that $M \models \Omega[\omega_1^N]$.

[15] Certainly if $M$ were extended to a new transitive model $N$ which contained these bijections, then the new $N$ would recognize these bijections for what they are.

[16] The details of this construction are too complicated to explain fully here. The relevant facts about $N'$ are these: 1.) $N'$ is a countable, transitive model of ZFC, 2.) $N'$ is an "end extension" of $N$ (i.e., for any $n \in N$, $\{x \mid x \in_{N'} n\} = \{x \mid x \in_N n\}$), and 3.) $N' \models$ "$\omega_1^N$ is countable". Further details about this type of construction can be found in chapter 7 of [30] or chapter 3 of [26]. For the purposes of this thesis, only the three above-listed facts about $N'$ will actually be used.

[17] The first of these equivalences follows simply from the fact that $N'$ is transitive. The second follows from the fact that

What's more, there's no possibility of explaining this divergence by appealing to the different ways $\Omega_M(\omega_1^N)$ and $\Omega_{N'}(\omega_1^N)$ interpret their quantifiers. Since $M$ and $N'$ have the same domain, $\Omega_M(x)$ and $\Omega_{N'}(x)$ interpret their quantifiers in exactly the same way. Hence, the difference in truth-value between $\Omega_M(\omega_1^N)$ and $\Omega_{N'}(\omega_1^N)$ *must* be explained in terms of the different ways $M$ and $N'$ interpret the symbol "$\in$." Indeed, following the analysis on pages 40–42, we can show that the difference between $\Omega_M(\omega_1^N)$ and $\Omega_{N'}(\omega_1^N)$ must be explained by the differing ways these formulas interpret *three particular* instances of "$\in$"—namely the three instances which are highlighted in the "$\in f$" clauses in the formulation of $\Psi(f, \omega_1^N)$.[18]

This example shows that there are versions of Skolem's Paradox—or, at least, of puzzles very much like Skolem's Paradox—which can only be resolved by appealing to the ways different models interpret the membership symbol. The example, of course, differs from Skolem's Paradox itself in that the example involves a comparison between two models, while Skolem's Paradox involves comparing one model to the entire set-theoretic universe. It's worth asking, therefore, whether the involvement of the whole set-theoretic universe makes a real difference to our analysis. To see why I think it doesn't make a genuine difference, it's useful to consider (very briefly) two more examples.

First, let's assume that the continuum hypothesis is true, and let's let $N$ be a countable, transitive model for ZFC. Using techniques pioneered by Barwise, we build a new model $N'$ such that 1.) $N \prec N'$, 2.) $N'$ is an *end extension* of $N$ (i.e., for every $n \in N$, $\{x \mid x \in_N n\} = \{x \mid x \in_{N'} n\}$), and 3.) $|N'| = \omega_1$.[19] Next, we let $\hat{m}$ be an arbitrary element of $N$ such that $N \models \Omega_N(\hat{m})$, and we let $X = \{g : \hat{m} \to \omega \mid g \text{ is a bijection}\}$. Finally, using the fact that $|X| = \omega_1$ (by the continuum hypothesis), we build a bijection $\sigma : N' \to N' \cup X$ such that $\sigma \restriction N = \text{Id}$; we then let $M$ be the model naturally induced by this $\sigma$—i.e., induced in the manner of footnote 6.

At the end of this construction, we have a model $M$ and an element $\hat{m} \in M$ such that three things hold: 1.) $\hat{m} = \{x \,|x \in_M \hat{m}\}$ is really countable, 2.) $M \models \Omega[\hat{m}]$, and 3.) $M$ contains *all* of the functions which witness the fact that $\Omega_E(\hat{m})$ is false—i.e., all of the $f$'s such that $\Psi_E(f, \hat{m})$ is true. Clearly, number 3 rules out the possibility that the difference between $\Omega_M(\hat{m})$ and $\Omega_E(\hat{m})$ can be explained by differing interpretations of the *initial* quantifier in $\Omega(x)$. Nor can we explain $M$'s failure to "recognize" instances of $\Psi_E(f, \hat{m})$ by noting that the quantifiers in $\Psi_E(f, \hat{m})$ range over a larger domain than those in $\Psi_M(f, \hat{m})$. After all, every

$N$ is transitive together with the above-mentioned fact that $\sigma \restriction X = \text{Id}$. Basically, the fact that $\sigma$ is the identity in the "neighborhood" of $\omega_1^N$ ensures that "locally-definable" properties of $\omega_1^N$ will be absolute between $M$ and $N'$ (and, for that matter, between either of these models and $V$). With a little work, we can show that *all* of the properties listed on page 41 are absolute once they have been reformulated to hold between $M$ and $N'$—i.e., rather than between (our original) $M$ and $V$.

[18] Very roughly, the fact that $\sigma \restriction X = \text{Id}$ ensures the vast majority of the symbols in $\Omega(\omega_1^N)$ occur within subformulas that are absolute between $M$ and $N'$. This fact, together with the fact that $\Omega_M(x)$ and $\Omega_{N'}(x)$ agree on the interpretation of quantifiers, is what allows us to isolate the relevant instances of "$\in$."

[19] The construction here rests upon the following theorem: *if $N$ is a countable model for ZF, then $N$ has an countable, elementary, end-extension.* To get an uncountable end-extension for $N$, we simply apply this theorem $\omega_1$ times, using union-of-chains arguments to get past limit ordinals. Since the properties of "being an elementary extension" and "being an end-extension" are both preserved through unions of chains, this construction gives us the desired $N'$.

set which is relevant to the *truth* of formulas like $\Psi_E(f, \hat{m})$—i.e., members of $\hat{m}$, $\omega$, $\hat{m} \times \omega$, etc.—*is* within the range of the quantifiers in $\Psi_M(f, \hat{m})$, and *every* set in the range of the quantifiers in $\Psi_M(f, \hat{m})$ is in the range of the quantifiers in $\Psi_E(f, \hat{m})$.[20] Hence, the *only* viable explanation for the difference in truth-value between $\Omega_E(\hat{m})$ and $\Omega_M(\hat{m})$ stems from the way these formulas interpret the symbol "$\in$."[21]

To give one final example of this kind of phenomenon—one in which the whole set-theoretic universe is still involved but in which there is *no* difference between the quantifiers in $\Omega_E(x)$ and those in $\Omega_M(x)$—we need to make a few more set-theoretic assumptions. Suppose, then, that $\omega_1^L$ is countable and that there is some definable bijective map, $F$, between $L$ and $V$.[22] In this case, we can assume without loss of generality that $F$ is actually the identity on some large initial segment of $L$—say, $L_{\omega_\omega}$—and we can use $F$ to define a new membership relation, $\in^*$, on $V$ such that $F$ becomes an "isomorphism" between $\langle L, \in\rangle$ and $\langle V, \in^*\rangle$. Then, everything I said about the penultimate case applies to this case as well. Even though $\langle V, \in^*\rangle$ and $\langle V, \in\rangle$ agree regarding the membership relation on $\omega_1^L$, $\Omega_{E^*}(\omega_1^L)$ is true and $\Omega_E(\omega_1^L)$ is false.[23] Further, because both of these sentences take their quantifiers to range over the entire set theoretic universe, the only possible explanation for this difference involves the sentences' divergent interpretations of "$\in$." We have, in short, a close relative of Skolem's Paradox in which the the (entire) set-theoretic universe plays a crucial role, but whose "solution" does not involve reflection on the ways different sentences interpret their quantifiers.

This, therefore, gives us a response to the first concern about the explanation of Skolem's Paradox with which this section began. Although it's quite true that the model $M$, "misses" a large number of bijections between $\hat{m}$ and $\omega$, this fact doesn't *explain* the failure of $(\dagger'_M)$. For one thing, it's hard to see how additional bijections *could explain* the failure of $(\dagger'_M)$, given that there are no reasons for thinking that $M$ could recognize these bijections as bijections. For another, the phenomenon at issue in this instance of Skolem's Paradox looks remarkably like the phenomenon at issue in other puzzles which live in the near vicinity of Skolem's Paradox—e.g., puzzles which arise when we formulate $(\dagger')$ so as to hold between two models, or between an *uncountable* model and the whole set-theoretic universe, or between two different "versions" of the universe itself. With respect to none of these other puzzles, can we explain the failure of $(\dagger')$ by appealing to divergent interpretations of the existential quantifier. Hence, we should eschew this explanation in the

[20] As in our original example, this point can be put in terms of "bounding sets." For any particular $f : \hat{m} \to \omega$, all of the quantifiers in $\Psi(f, \hat{m})$ are bounded by sets like $f$, $\hat{m}$, $\omega$, $\hat{m} \times \omega$, etc. Since all of the elements of such sets are in $M$, the quantifiers in $\Psi_M(f, \hat{m})$ "know" about these elements.

[21] As with each of the last few examples, a careful analysis of this example allows us to isolate the difference between $\Omega_E(\hat{m})$ and $\Omega_M(\hat{m})$ in the different interpretations these sentences give to the three instances of "$\in$" which occur in the "$\in f$"-clauses in $\Psi(f, \hat{m})$. Since there's nothing new going on in this particular case, I omit the details of this analysis here.

[22] How strong are these assumptions? From the standpoint of consistency strength, not very. Given any transitive model of ZFC, we can build a new model, $M$, such that $M \models \text{ZFC} + |\omega_1^L| = \aleph_0$ and such that there is a definable bijection between $M$ and $L \cap M$. Hence, the assumptions above are equaconsistent with ZFC.

That being said, these assumptions *do* go beyond what ZFC can actually *prove* (so they may be false in the real set-theoretic universe). Given the purposes of the present example, I don't think this possibility matters too much; it should, however, be kept in mind.

[23] Here $\Omega_{E^*}(\omega_1^L)$ is the obvious modification of $\Omega_E(\omega_1^L)$ in which our semantics interpret "$x \in y$" as meaning $F^{-1}(x) \in F^{-1}(y)$.

present case as well. This is especially true, given that we have *another* explanation—i.e., the explanation in terms of "$\in$"—which *does* work in a uniform manner across all four examples.

This brings us to a second concern. Recall how the model $M$ was originally constructed. We began with a transitive model, $N$, and we modified $N$ by "replacing" one of its members with the function $f$; at the end of this construction, $N$ and $M$ wound up being isomorphic, and we used this fact in showing that $\Omega_M(\hat{m})$ was true. How, then, can we have arrived at a different analysis of Skolem's Paradox in this case than we did in the transitive-models case discussed in section 2.1? If $M$ and $N$ are isomorphic, shouldn't our explanation of the difference between $\Omega_E(x)$ and $\Omega_M(x)$ be exactly the same—or, at least, structurally similar—to our explanation of the difference between $\Omega_E(x)$ and $\Omega_N(x)$? How can the one explanation involve (only) the interpretation of "$\in$," while the other involves (only) the interpretation of quantifiers?

The short response to these questions is that the properties preserved under isomorphisms comprise a (far) smaller class than do the properties which semantics—of either the ordinary English or the model-theoretic variety—pays attention to. So, even though isomorphisms preserve the overall size of a model, they do not preserve the *specific elements* which make up a model's domain. In contrast, the quantifiers in both $\Omega_E(x)$ and $\Omega_M(x)$ *do* look at specific elements—e.g., the quantifiers in $\Omega_M(x)$ range over (exactly) the elements in $M$ and not over the elements in other, equivalently *sized,* domains. Similarly, although two isomorphic models have to share the same overall "pattern" of membership relations—in the sense that their relations get "lined up" with each other via the relevant isomorphisms—they need not agree on any specific questions about membership—i.e., for $N \simeq N'$ and $n, n' \in N \cap N'$, there is no requirement that $n \in_N n' \Leftrightarrow n \in_{N'} n'$. In contrast, the semantics of $\Omega_E(x)$ and $\Omega_M(x)$ involve *specific* understandings of the membership relation.

This gives us an abstract answer to our question: because semantics deals with more fine-grained notions than isomorphisms can capture, we should not be surprised to find that isomorphic models give rise to very different versions of Skolem's Paradox. To make this abstract answer more concrete—and hopefully more convincing—it's worth examining one non-set-theoretic case in which a similar phenomenon arises. Consider, therefore, the following three models in the language containing only a single unary function symbol, $s$:[24]

- Domain($N_1$) = $\mathbb{N}$; $s_1 = \{\langle n, n+1 \mid n \in N_1\rangle\}$.
- Domain($N_2$) = $\mathbb{N} \setminus \{0\}$; $s_1 = \{\langle n, n+1 \mid n \in N_2\rangle\}$.
- Domain($N_3$) = $\mathbb{N}$; $s_3 = \{\langle n, n+1 \mid n \in N_1 \ \&\ 2 < n\rangle\} \cup \{\langle 1, 0\rangle\} \cup \{\langle 0, 2\rangle\}$

Consider also the obvious way to define "Zero" in these models:

$$\text{Zero}(n) \equiv_{df} \neg\exists x\, [s(x) = n].$$

As noted in section 2.1 (cf. p. 36), $N_1 \models \text{Zero}(0)$ while $N_2 \models \neg\,\text{Zero}(0)$. Further, the only explanation for this difference stems from the different domains over which the initial quantifier in the definition of "Zero"

[24] This example is simply a modification of one given on page 36 For expository convenience—i.e., to avoid making the reader flip back-and-forth—I repeat most of the details of that example here.

is allowed to range. In contrast, $N_3$ also satisfies $\neg\,\mathrm{Zero}(0)$, but the difference between $N_1$ and $N_3$ cannot be explained by appealing to differing domains of quantification (since $N_1$ and $N_3$ have exactly the *same* domain). Instead, this latter difference *must* be explained by the different ways $N_1$ and $N_3$ interpret the function symbol, $s$.[25]

For our purposes, the important point about this example is this: even though we get different explanations for the divergence between $Zero_{N_1}(0)$ and $Zero_{N_2}(0)$ and the divergence between $Zero_{N_1}(0)$ and $Zero_{N_3}(0)$, the models $N_2$ and $N_3$ are clearly isomorphic. What's more, both $N_2$ and $N_3$ are also isomorphic to $N_1$. Hence, mere isomorphism is not enough to ensure that two models share the same semantics. Nor is an isomorphism between two models enough to ensure that semantic differences between these models and some third model (or ordinary English, for that matter) must have the same "explanation." For the purposes of semantics, then, isomorphism is mostly a red herring.

This provides a complete answer to our second concern. Even though $M$ and $N'$ are isomorphic, this gives no reason for thinking that the difference between $\Omega_M(x)$ and $\Omega_E(x)$ will run parallel to that between $\Omega_{N'}(x)$ and $\Omega_E(x)$. Hence, it gives no reason for thinking that there's something problematic about an explanation of Skolem's Paradox which focuses on the interpretation of "$\in$" rather than the interpretation of "$\exists$." Although this explanation is different from the "usual" explanation of Skolem's Paradox, it's clearly the right explanation for the present case.

## 2.4 Conclusion

In sections 1.1–1.2, I formulated a fairly simple version of Skolem's Paradox and showed how that version of Skolem's Paradox depends on a particular claim about the relationship between model theory and ordinary mathematical English—i.e., ($\dagger'$). In section 1.3, I showed that (and why) this claim has to be false. The explanation I gave there suffered, as noted on pages 28–30, from three defects. First, it addressed only the simple version of Skolem's Paradox developed in 1.1–1.2 and ignored the more sophisticated versions of Skolem's Paradox which involve, for example, transitive models and/or elementary submodels of the universe. Second, although it isolated several semantic differences between $\Omega_E(x)$ and $\Omega_M(x)$, and although it showed that *one of* these differences had to be responsible for the failure of ($\dagger'$), it didn't pin down the specific difference which actually does the explanatory work. Finally, my whole argument assumed that the reader was simply *puzzled* about Skolem's Paradox—i.e., that they found the paradox perplexing, but had no *theoretical* reasons for thinking that the semantics of $\Omega_E(x)$ ought to be identified with the semantics of $\Omega_E(x)$. In particular, I made no attempt to respond to any of the arguments which might be made on behalf of principles like ($\dagger'$).

[25] It's worth noting here that, because of the way $N_1$ has been defined, the semantics induced by $N_1$ are "essentially" the same as those of ordinary arithmetical English. In particular, for any formula $\Phi(\bar{x})$ in our language, and any $\bar{n}$ in $\mathbb{N}^n$, we get the equivalence: $\Phi_E(\bar{n}) \iff \Phi_{N_1}(\bar{n})$. Hence, everything I say about the differences between $\mathrm{Zero}_{N_1}(x)$ and $\mathrm{Zero}_{N_2}(x)$ (or $\mathrm{Zero}_{N_1}(x)$ and $\mathrm{Zero}_{N_3}(x)$) holds also about $\mathrm{Zero}_E(x)$ and $\mathrm{Zero}_{N_2}(x)$ (or $\mathrm{Zero}_E(x)$ and $\mathrm{Zero}_{N_3}(x)$).

The present chapter has tried to deal with the first two defects. On the one hand, I have canvassed a number of fairly sophisticated formulations of Skolem's Paradox and have argued that they all fail for essentially the same reasons as the original (simple) formulation failed. In each case, there was some difference in the way $\Omega_E(x)$ and $\Omega_M(x)$ interpret the symbols "$\in$" and "$\exists$" which explains—and explains in a fairly intuitive way—why the relevant version of ($\dagger'_M$) has to fail. On the other hand, I have argued that the request for a uniform explanation as to *which* symbol causes the failure of ($\dagger'_M$) is misguided. For different choices of $M$, different symbols will play the crucial explanatory role. Hence, unless we have quite a bit of information about the particular model in terms of which Skolem's Paradox has been formulated, we can't say much more than I said in my original discussion in 1.3.

Of course, none of this deals with the third defect. What's more, insofar as I have missed some obvious arguments in favor of ($\dagger'$), my analysis of even the sophisticated cases of Skolem's Paradox will continue to be incomplete. In order to remedy this third defect, I turn in chapter 3 to examine some of the arguments which might be made on behalf of ($\dagger'$).

# Chapter 3

# Two Philosophical Objections

In this chapter, I consider two objections to the (essentially technical) solution to Skolem's Paradox which I have developed over the last two chapters. Although I don't, in the long run, think that either of these objections is convincing, I do think they uncover much of the philosophical allure of Skolem's Paradox and thereby help to explain the philosophical community's recurrent flirtation with it.[1] I also think these objections raise independent questions concerning the broader philosophical significance of modern model theory. Hence, whether or not they ultimately support Skolem's Paradox, these objections are rich enough to repay a fair bit of close philosophical attention.

## 3.1 Naive Prattle

One objection which could be raised to my solution of Skolem's Paradox involves the use this solution makes of the account of $\Omega_E(x)$ developed in section 1.2.1. There, I said that the semantics of $\Omega_E(x)$ can be identified with those of the sentence Cantor(x), where the semantics of Cantor(x) are those of "ordinary mathematical English." I then explained—in 1.3 and 2.1–2.3—how these "ordinary-English" semantics differ from the model-theoretic semantics of $\Omega_M(x)$ and how this difference, in turn, explains what's going on in Skolem's Paradox (and, in particular, why the conditional $(\dagger'_M)$ fails).

This whole argument, however, assumes that sentences like Cantor(x) *have* a determinate semantics. So, for instance, it assumes that the phrase "$x$ is a member of $y$" lets us refer to the "real" membership relation on the set-theoretic universe. Similarly, it assumes that the phrase "there is a set $x$ such that..." allows us to evaluate "..." at *every* set in the set-theoretic universe. Without these assumptions, my comparison between $\Omega_E(x)$ and $\Omega_M(x)$ comes apart, and my overall explanation of the failure of $(\dagger'_M)$ becomes ungrounded.

These reflections give rise to an objection to the analysis of chapters 1–2 which I like to call the "naive

[1] As noted in the introduction, there are really two puzzles associated with Skolem's Paradox. First, there is the paradox itself: how can a countable model satisfy the very axioms which prove that uncountable sets exist? Second, there is a problem concerning the *reception* of Skolem's Paradox: why, even though the technical "solution" to Skolem's Paradox has been pretty well-understood since the early twenties, have philosophers continued to find this paradox so tempting? The last two chapters focused on the first of these puzzles; this chapter focuses on the second.

prattle objection." The objection begins by raising concerns about the determinacy of "ordinary mathematical English." It then argues that these concerns undercut the analysis of ($\dagger'_M$) which I have provided, and that, as a result, my overall solution to Skolem's Paradox is a failure. In short, the fact that my solution presupposes the intelligibility of ordinary mathematical English—of ordinary talk about sets and classes, membership and non-membership, countable sets and uncountable sets—renders this solution nothing more than "so much naive prattle."

Before discussing specific developments of this objection, I want to make four points about the objection's general structure. First, the objection doesn't challenge the purported *connection* between the semantics of $\Omega_E(x)$ and those of Cantor($x$): insofar as $\Omega_E(x)$ is simply an abbreviation of Cantor(x), the semantics of these two sentences must be the same. Instead, the objection argues that the semantics of Cantor(x) are indeterminate to begin with, and it then *uses* the connection between Cantor(x) and $\Omega_E(x)$ to show that this indeterminacy carries over to $\Omega_E(x)$.

Second, this objection threatens my analysis of Skolem's Paradox at several different levels. At the most obvious level, it undercuts my analysis of the semantic differences between $\Omega_E(x)$ and $\Omega_M(x)$ (since this analysis presupposed that we could use phrases like "all sets" or "is *really* a member of" in a naive manner when explaining the semantics of $\Omega_E(x)$). At a deeper level, the objection undercuts much of my preliminary discussion of Skolem's Paradox in sections 1.1–1.2. So, for instance, my insistence in 1.1 that we distinguish between $\{x \mid x \in \hat{m}\}$ and $\{x \mid x \in_M \hat{m}\}$ depends on the assumption that we can isolate the "real members" of $\hat{m}$. Similarly, my argument for reading premise 2 in terms of $\Omega_E(x)$—the argument given in 1.2.1 and 1.2.2—assumes that phrases like "$x$ is uncountable" and "$x$ is a member of $y$" are themselves determinate. Finally, and most importantly, my initial discussion of premise 5—the discussion in section 1.1 which led to the conclusion that this premise is true under "interpretation I"—depends on the assumption that we can use the phrase "is countable" with enough precision to make determinate claims about the "countability" of models and their subsets.

Third, and as a direct consequence of the second point, the current objection shouldn't be viewed as a *defense* of ($\dagger'_M$). Because the objection undercuts *both* my analysis of ($\dagger'_M$) *and* my argument for regarding ($\dagger'_M$) as the key to Skolem's Paradox, a proponent of this objection could reject ($\dagger'_M$) while still defending Skolem's Paradox. He might, for instance, claim 1.) that ($\dagger'_M$) is not the heart of Skolem's Paradox, 2.) that ($\dagger'_M$) is meaningless, and 3.) that the version of Skolem's Paradox introduced at the beginning of section 1.1 is valid. In this case, the current objection would serve to bolster number 1, while 1 itself would serve to render 2 and 3 compatible.

That being said, this objection clearly *makes room* for understandings of set theory on which ($\dagger'_M$) comes out true. So, for instance, (purported) difficulties with the "ordinary-English" semantics for $\Omega_E(x)$ might lead us to "clarify" this sentence by *identifying* its semantics—and, by extension, those of Cantor(x) and "$x$ is uncountable"—with the semantics of $\Omega_M(x)$. These new semantics would, of course, be "relative" to the particular $M$ which is under discussion, but they would also validate ($\dagger'_M$). Further, it's relatively easy to see how such semantics could be used to formulate an interesting version of Skolem's Paradox.[2]

Finally, this objection isn't specifically aimed at *my* solution to Skolem's Paradox. Instead, it raises problems for *any* solution to Skolem's Paradox which takes classical set theory at face value and then uses this set theory to undermine Skolem's Paradox. So, for instance, even a solution which attributes all failures of ($\dagger'_M$) to the interpretation of the initial quantifiers in $\Omega_E(x)$ and $\Omega_M(x)$—i.e., one which resists the arguments of 2.2 and 2.3—will still be subject to this objection. Hence, whether or not the objection can ultimately be parlayed into a substantial *formulation* of Skolem's Paradox, its scope and generality make it an effective *counter* to most "standard" solutions to that paradox.

### 3.1.1 Three Examples

Why might someone think that the semantics of $\Omega_E(x)$ are too indeterminate to support the argument of the last two chapters? As far as I can see, there are three lines of argument which might be used to support this view. The first, and most obvious, stems from simple incredulity that anything as strong as classical set theory could simply be *presupposed* when solving philosophical problems. Surely there is something problematic about *just assuming* the existence of all the denizens of the set-theoretic universe—of uncountable many subsets of $\mathbb{N}$, of cardinal numbers which are so large that they can't be described in ordinary English, and of rank-upon-rank of uncountable sets stretching beyond the point where (even) set theory can count the ranks themselves. Surely there is something even more problematic about assuming that we can *talk about* this whole menagerie in a determinate manner. At the very least, these assumptions seem to demand substantial philosophical arguments before they can be used to "solve" Skolem's Paradox.

In the literature, this concern about assuming too much when solving Skolem's Paradox tends to get phrased in terms of "Platonism." In [58], for instance, William Thomas argues that it is "the very essence of Platonism" to assume that ordinary language can be used as "a kind of universal metalanguage" when discussing Skolem's Paradox; he also argues that we display "creeping Platonism" when we attempt to talk about "absolutely uncountable sets" like "the set of *all* real numbers" (see [58], p. 195 and 196). Similarly, in [18], Arthur Fine argues that, "unless one is to fall back on some strong form of Platonism," one must accept that the phrase "all real numbers" is systematically ambiguous—i.e., since it, or at least its formal analogue, can be interpreted at various non-isomorphic models (see [18], p. 30). Finally, in [28], Virginia Klenk characterizes as "Platonists" those who "appeal to the distinction between formal and informal, or intuitive, [or ordinary-English] mathematics" (see [28], p. 480–481). She then argues that Platonism is far too problematic a position to ground a satisfactory solution to Skolem's Paradox.

So far, of course, we have merely put a label on the assumption that $\Omega_E(x)$ is semantically determinate

[2] This seems to be the line taken by Fine in [18] and by Thomas in [59]. In both cases it is claimed that the only sense we can make of phrases like "$x$ is uncountable" comes from interpreting these phrases *at a particular model*—i.e., from identifying their semantics with those of sentences like $\Omega_M(x)$ for a particular choice of $M$. Since the model under discussion varies according to context, the phrase "$x$ is uncountable" winds up being intrinsically relative.

and have hoped that this label is sufficiently pejorative to make people question that assumption. To get a real *argument* against the assumption, we need to turn to a second reason for questioning the semantic determinacy of $\Omega_E(x)$. This reason stems from reflection on some fairly *general* concerns about the determinacy of mathematical English—i.e., on concerns which are not specific to the case of $\Omega_E(x)$ and which may not even be specific to the case of set-theory. We might, for instance, appeal to the concerns Benacerraf raises in [2] or [3]. Or, we might appeal to the concerns raised by Field in [16] or Kitcher in [27]. In any of these cases, general arguments against the determinacy of mathematical language are imported into the discussion of Skolem's Paradox and parlayed into specific arguments against the determinacy of $\Omega_E(x)$.

This "importation" strategy is actually fairly common in the literature on Skolem's Paradox. In [28], for instance, Virginia Klenk presents two arguments against "anti-skolemite" appeals to "informal models of set theory" and/or "ordinary language conceptions of sets."[3] First, she argues that solving Skolem's Paradox by appealing to "informally specified models" reverses the real relationship between formal and informal mathematics, between model-theoretic semantics and ordinary mathematical English. On Klenk's account,

> Our intuitive mathematics should be seen as something less, not more, than our formal mathematics: less precise, less consistent, perhaps, and less complete (unless, of course, it is inconsistent, in which case its completeness hardly counts as an advantage)." ([28], p. 482)

On this view, then, there is no reason to think that informal mathematics—or, in our terms "ordinary-English" mathematics—has any special ability to pick out the "real" notion of set or the "real" conception of countability. To the extent that formal axioms fail to "capture" or "specify" these notions—e.g., in the ways highlighted by Skolem's Paradox—the notions are simply unspecifiable:

> Intuitive mathematics must, after all, be couched in language, and there is little reason to think that the ordinary language of informal mathematics is in any way superior to the formal language of first-order predicate logic . . . Once we have formalized the notion, moved from ordinary language to the formal language, what *more* could there be to the informal theory, unless one supposes that vagueness is a hallmark of comprehensiveness. Where is this "real" concept of set to be found? ([28], p. 482)

At the very least, Klenk argues, those who want to maintain that informal mathematics *can* do more than formal mathematics, ought to have a well-developed account of *how* "ordinary English" gains its special powers. In the absence of such an account, the retreat to ordinary English seems counterintuitive.

Klenk's second argument against the use of ordinary English in set theory comes from a concern about multiple realizations of the set-theoretic universe. To make this concern perspicuous, Klenk begins by

[3] A bit of background may be helpful here. In [47], Michael Resnik argues that informal mathematics provides a perfectly good interpretation of the notion "all sets," and that this interpretation provides us, in effect, with a model for the language of set theory (the "informal model" mentioned above). Because Skolem's Paradox only arises when we ignore this model and try to interpret set-theoretic language at other (specifically, countable) models, the decision to treat this "informally specified" model as canonical—i.e., canonical for interpreting phrases like "all sets" or "x is uncountable"—allows us to solve Skolem's Paradox. In particular, it explains why we should resist the urge to infer "x is uncountable" from mere fact that some (non-canonical) model satisfies the formal analogue of "x is uncountable."

Klenk's paper raises doubts about Resnik's invocation of "informally specified" models of set theory. In the long run, Klenk thinks that reflection on Skolem's Paradox should push us towards some kind of formalism.

examining a similar problem in the context of number theory. Essentially, she notes that a simple application of the trick from footnote 6 allows us to build an arbitrarily large collection of models all of which are isomorphic to the "standard" model of number theory—i.e., to $\mathbb{N}$—but which disagree about the actual ontology of number theory:[4]

> . . . there are many different sorts of things which could serve as a basis for standard models. We might take numbers just as indefinable abstract entities, or we could construe them as sets. In the latter case, of course, there are various possibilities: we could take the number two, for instance, either as $\{\{0\}\}$ or as $\{0, \{0\}\}$. We might even take numbers as collections of inscriptions.

Since all of these models are isomorphic, Klenk sees no reason to privilege *one* of them as the "real" sequence of natural numbers. Instead, she suggests that any of these "number sequences" can *count as* the natural numbers. As a result, any particular object can *count as* any particular number, so long as it appears in the right place in some appropriate "number sequence."[5]

Next, Klenk notes that a similar argument can be mounted in the case of set theory. Even if we ignore the queer models produced by Löwenheim-Skolem arguments, it's easy to generate nice models using the same isomorphism tricks deployed in the number-theoretic case. If, for instance, there is something which lives outside the set-theoretic universe—say, my cat Gandalf—then we can consider a correspondence which leaves all sets other than $\aleph_0$ fixed while swapping $\aleph_0$ for Gandalf. Similarly, we can take any two sets—say, $\aleph_0$ and $\aleph_1$—and consider a correspondence which permutes these two sets while leaving everything else fixed. In either case, we generate an interpretation of set theory which is structurally indistinguishable from the one with which we started.[6] By parity of reasoning—i.e., parity with the reasoning used in the number-theoretic case—Klenk concludes that all of these interpretations should count as "intended interpretations" for the language of set theory.[7]

This conclusion, however, seriously undercuts the kind of analysis of Skolem's Paradox which I gave in chapters 1 and 2. My analysis assumed that there is a determinate significance to the English phrase "all sets." It assumed, that is, that there is a unique answer to the question "which things count as sets and which do not?" Now, however, it turns out that *any* object counts as a set under some (intended) interpretations

[4] The quote here is from ([28], p. 484). The basic technical point is simple: since any countably infinite set can be put into bijective correspondence with the natural numbers, we can obtain a model for the language of number theory which 1.) uses this set as its domain and 2.) is isomorphic to the "real" natural numbers. Further, since any particular element of our set can correspond to any particular natural number, any element can "serve as" any natural number in the resulting model.

[5] For a more detailed development of this position, see [3]. It should be noted that this position—commonly dubbed "structuralism"—has recently become quite popular among philosophers of mathematics. See, for instance, [23] or [49].

[6] A technical comment is in order here. Since the correspondences under consideration permute the *entire* set-theoretic universe, they cannot properly be regarded as *isomorphisms* (since they are not even sets). Hence the theorem discussed in footnote 6 does not strictly apply. However, the correspondences in question are so simple that it is trivial to define directly the interpretations of "$\in$" and "all sets" which these correspondences naturally induce. Hence, the fact that our argument is not a strict application of the theorem from footnote 6 doesn't really matter.

[7] Again, this kind of position has obtained a good bit of popularity in the recent literature. For an extended discussion (and defense) of the position, see [23].

of set theory and that *no* object counts as a set in all such interpretations. In a similar manner, my analysis assumed that the word "membership" has a specific "ordinary-English" meaning, and that this meaning is enough to determine when one thing is *really* a member of another thing. Now, however, it seems as though *any* object can be a "member" of *any* other object, provided we use the right "intended interpretation" of set theory. Given all this, the determinacy of "ordinary mathematical English" begins to seem far too fragile to support the arguments of chapters 1 and 2.

These, then, are Klenk's two arguments against the naive invocation of "ordinary-English" interpretations of phrases like "set" and "membership." Leaving these arguments aside, let me briefly mention two other objections to such invocation. In [59], William Thomas notes that a philosopher who *starts out* with philosophical commitments that are incompatible with traditional set theory will be unimpressed by arguments which assume that set theoretic language has a determinate "ordinary-English" semantics. So, for instance, a formalist will be unimpressed, because he doesn't think that *any* mathematical language has significance over-and-above its role in formal proofs (certainly he doesn't think that set-theoretic language *refers* to a real set-theoretic universe).[8] Similarly, an adherent of Wang's system of set theory will reject such arguments because his own positive understanding of set theory entails that there is no "absolute" significance to phrases like "all sets" or "is uncountable."[9] In both cases, then, the fact that a philosopher has positive views which tell against the determinacy of "ordinary-English" set theory—at least when this set theory is assumed to bear a close resemblance to Cantor's set theory or to informal versions of ZFC—provides this philosopher with ample reasons for rejecting arguments which use ordinary-English set theory to "solve" Skolem's Paradox.

A somewhat different objection to ordinary-English set theory has been raised by Crispin Wright. Following Michael Dummett, Wright argues that the significance of terms in ordinary language must be manifested in the ways we use this language—that, to cite Wittgenstein, "meaning cannot transcend use." He then argues that it's hard to see how our use of set-theoretic language (whether formal or informal) could possibly be rich enough to fix a unique interpretation for that language. What would be the practical differences

[8] See [59] pp. 180–181 for this argument about formalism. See also the last section of [28].

[9] In Wang's system, sets are stratified into an ascending sequence of levels, $\langle \Sigma_\alpha \mid \alpha < \delta \rangle$, where $\delta$ is an (unspecified) countable ordinal. (In practice, each $\Sigma_\alpha$ can be identified with $L_{\alpha+\omega}$ in the ordinary stratification of $L$.) The system allows quantification over particular levels—using explicitly indexed quantifiers $\forall_\alpha$ and $\exists_\alpha$—but it does not allow quantification over the totality of all levels. In addition, although Wang's system allows that certain sets are uncountable *in* $\Sigma_\alpha$—i.e., that there are no enumerating functions for these sets in $\Sigma_\alpha$—it also proves that these sets are countable *in* $\Sigma_{\alpha+2}$—i.e., that there *are* enumerating functions at this higher level.

Thomas' point, then, is just that an adherent of Wang's system is likely to be suspicious of informal appeals to "all sets," since his own set theory explicitly forbids such unrestricted quantification. Indeed, as Thomas notes, a proponent of Wang's system is likely to balk at (even) notions like "all real numbers" (since the reals in Wang's system are spread *throughout* the hierarchy of $\Sigma_\alpha$'s). Similarly, because the countability of particular sets can change depending on the "level" from which you look at it, and since *every* set winds up being countable at some level, informal appeals to "sets which are *really* uncountable" are likely to fall on deaf ears.

For more on Wang's system, see [62] or [63]. For Thomas application of Wang's system to Skolem's Paradox, see [58].

between people who were talking about the whole set-theoretic universe and people who were talking about some (elementarily equivalent) submodel of the universe?[10] Unless we have a reasonable answer to this question—or at least an idea of how such an answer might start to go—we should be cautious about assuming that there *is* any "ordinary-English" understanding of set theory (or, at least, any understanding which goes beyond the insistence that we refrain from using set-theoretic language in ways which violate the axioms of ZFC). In particular, we should avoid falling back on naive appeals to the "ordinary-English" understanding of $\Omega_E(x)$ when we set about solving Skolem's Paradox.[11]

These, then, are four reasons for questioning the semantic determinacy of $\Omega_E(x)$. The assumption that $\Omega_E(x)$ has a determinate semantics seems to misconstrue the relationship between formal and informal mathematics (Klenk); it ignores difficulties stemming from multiple realizations of the set-theoretic universe (Klenk again); it can be resisted by philosophers whose underlying commitments tell against ZFC style set theory (Thomas); and it runs afoul of Wittgenstein's dictum that "meaning cannot transcend use" (Wright). Clearly there are other arguments which could be developed here. Equally clearly, the arguments above could be developed in (far) more detail. For the present, however, I eschew such development and simply take the sketches above to illustrative some of the ways in which general concerns about the determinacy of mathematical language can be imported into discussions of Skolem's Paradox and used to undercut the solution to that paradox which I have developed.

At this point, then, I turn to a final argument for questioning the determinacy of sentences like $\Omega_E(x)$. Unlike the arguments above, this one does not involve any general skepticism about mathematical language. Instead, it involves some fairly specific worries about *set-theoretic* language, worries which arises from the rather peculiar history of this discipline. Three features of this history warrant special attention here.

First, set theory is often regarded as a discipline "born in the sin of contradiction." The history of set theory is full of cases where the naive use of set-theoretic reasoning led to contradictions, with the

[10] There are two questions here which should be kept distinct (and which Wright takes great pains to distinguish in his paper). One concerns whether we, as outsiders, could ever *tell* whether people were talking about the whole universe or only an elementarily equivalent submodel. The other concerns whether there is anything *about the people's practices*—i.e., about the things they say and don't say, about their underlying dispositions to say things in the future, about the particular discriminations they make (or are inclined to make) in particular cases, etc., etc.—which fixes what they're talking about. Wright's paper is concerned only with the latter, and far more difficult, question.

[11] Wright's argument here can be found in [65] pp. 126–137. It's worth noticing that this Wittgenstinian argument is only one of several which Wright gives in this paper. So, for instance, Wright also develops a variant of Klenk's argument concerning the relationship between formal and informal mathematics. After arguing that axioms provide explications of mathematical concepts, he writes:

> ... a good explication cannot be *weaker* than the intuitive concept is supplants, cannot be neutral on points on which the latter is committed. Accordingly if the ZF-axioms, with '$\in$' interpreted as set membership, did constitute a satisfactory explication of the intuitive notion of set, the fact that they do not, so interpreted, entail the existence of uncountable sets would force the conclusion that there is no such entailment from the intuitive concept of set either. ([65] p. 126)

Wright also follows Klenk in providing an extended discussion of the relationship between Skolem's Paradox and the problem of multiple realizations of number theory (see pp. 120–124).

most famous example being the discovery that unrestricted versions of the comprehension principle lead to Russell's Paradox.[12] Even leaving this paradox aside, it's relatively easy to find other instances where naive reasoning about sets goes wrong. One need only mention Konig's Paradox, Richard's Paradox and the Burali-Forti Paradox.[13] This tendency towards contradiction, then, might lead us to be suspicious of solutions to Skolem's Paradox which take ordinary-English talk about "all sets" (or about sets which are "really uncountable") at face value.[14]

Second, the response to these early contradictions often involved the establishment of rival, and in many cases incompatible, versions of set theory. Russell himself dealt with his paradox by adopting the theory of types.[15] Other set theorists turned to Zermelo's axiomatization of set theory, while still others turned to axiomatizations by, e.g., Quine and Wang.[16] Insofar as each of these theories is intended as an elaboration of our initial (naive) understanding of set theory, the fact that they wind up in such different places—both with respect to their underlying intuitive motivations and their actual technical results—might lead us to think that there *isn't* a determinate "ordinary-English" conception of sets from which they all spring.

Third, even if we discount the significance of non-standard versions of set theory and follow the majority of set theorists in accepting Zermelo's axiomatization as "canonical," the fact that this axiomatization leaves many *basic* questions about the set-theoretic universe undecided might raise worries about the de-

[12] Here, unrestricted comprehension says that for any property $P$ there is a set $X$ which contains exactly those things that possess property $P$: $\forall P \exists X \forall y\, [y \in X \leftrightarrow P(y)]$. Defining $P(y) \equiv y \notin y$, this principle produces a set $X$ such that $y \in X \leftrightarrow y \notin y$. But, this immediately leads to Russell's Paradox:

$$X \in X \leftrightarrow X \notin X.$$

From a historical perspective, there are two things to mention about this paradox. First, the paradox wasn't originally formulated in the context of set theory. Russell originally formulated the paradox in the context of Frege's theory of "concepts" and "extensions," and it was this formulation which forced Frege, in the long run, to abandon his theory. Second, the set-theoretic version of Russell's Paradox was well-known to Cantor and his early followers, and they had a number of tools for dealing with it. Hence, although the paradox has serious implications for an absolutely naive approach to set theory—i.e., one which uses an absolutely unrestricted version of comprehension—it didn't have much of effect on Cantor's own understanding of set theory. For more on Cantor's understanding of these issues, see [10], [20] or [35].

[13] See [50], [29] and [8]. It should be noted that, from a modern perspective, these three paradoxes are rather different in character. The first two are really paradoxes in semantics. The third is a genuine problem in set theory, as it's solution requires us to reformulate the basic principles of naive set theory to prevent the class of all ordinals from constituting a set.

[14] This is particularly true insofar as two of the paradoxes above—Konig's Paradox and Richard's Paradox—depend crucially on our tendency to use ordinary-English in set theory. (Essentially, they exploit this tendency to import semantic paradoxes about ordinary English *into* the realm of set theory.) Whatever we think about set theory in general, these paradoxes might lead us to think that "ordinary-English" set theory is deeply problematic.

[15] See [51] or [52].

[16] See [68], [66], [46], [62] and [63]. Note that these axiomatizations disagree with each other about important technical questions. Quine's set theory, for instance, proves the existence a universal set; the others prove that no such set exists. Similarly, Zermelo's set theory "stratifies" the set-theoretic universe into an uncountable sequence of levels; the theories of Russell and Wang also stratify the universe, but they only allow countably many levels in their sequences. Finally, Wang's set theory ensures that, for any two infinite sets $X$ and $Y$, there is a bijection $f : X \to Y$; each of the other set theories allows (indeed proves) the existence of "absolutely" uncountable sets.

terminacy of set-theoretic language. So, for instance, the fact that ZFC cannot determine whether all sets are "constructible" from simpler sets—whether, that is, the axiom $V = L$ is true—raises worries about our understanding of the notion "all sets."[17] So too does the fact that that ZFC cannot answer fundamental questions about the "height" of the set-theoretic universe.[18] Finally, the fact that ZFC leaves natural questions about cardinality open—e.g., what is the cardinality of $\mathcal{P}(\omega)$? is $\omega_1^L$ really uncountable?—raises concerns about the determinacy of "plain English" talk of cardinality (e.g., is there any "absolute" meaning of "uncountable" which will settle my question about the cardinality of $\omega_1^L$?).

These, then, are three "historical" reasons for questioning the determinacy of "ordinary mathematical English." The fact that naive set theory has often lead to contradictions makes us worry about the *coherence* of ordinary talk about sets; the fact that there are competing versions of set theory makes us worry about the *determinacy* of "plain English" appeals to "all sets" or "is really a member of"; the fact that (even) the standard axiomatization of set theory leaves *basic* issues in set theory undecided makes us worry about the *absoluteness* of notions like "all sets" or "is really uncountable." In light of all this, it seems hopelessly naive to think that we can solve Skolem's Paradox by just assuming that sentences like $\Omega_E(x)$ have a determinate semantics and then using a straightforward analysis of this semantics to undermine the paradox.

Before leaving these historical arguments behind, a final comment is in order. Because these arguments focus specifically on the history of *set theory*—as opposed to mathematics in general—they may help to explain why philosophers tend to be more impressed with Skolem's Paradox than with analogous puzzles about notions like "finite."[19] So, for instance, it's relatively easy to show that there is no collection of sentences $\Gamma$ such that for any model $M$,

$$M \models \Gamma \iff M \text{ is finite.}$$

Similarly, given any axiomatization of set theory, $\Gamma$, it's easy to show that there is no formula $\Omega_f(x)$ such that for any $M \models \Gamma$ and any $\hat{m} \in M$,

$$M \models \Omega_f(\hat{m}) \iff \{m \in M \mid M \models m \in \hat{m}\} \text{ is finite.}$$

Thus, to the extent that concerns about Skolem's Paradox are motivated by *general* arguments against the determinacy of mathematical language, these concerns should carry over from the case of "countable"—i.e., the case involved in Skolem's Paradox—to the case of "finite" as well.

[17] The problem here is not simply that some sentences can't be decided by ZFC. Rather, it's that the axiom $V = L$ raises a *conceptual* issue about set theory: what is the relationship between set formation and definition? It seems, therefore, to be the *kind* of issue on which axiomatizations of set theory should take a stand. Hence, the fact that ZFC cannot settle this issue—cannot, that is, decide whether the phrase "all sets" includes non-constructible sets—might make us wonder about our grasp on the notion "all sets."

[18] e.g., does the set-theoretic universe *contain* inaccessible cardinals, or does the class of ordinals, in effect, *constitute* the first inaccessible cardinal?

[19] See, for instance, Crispin Wright's remarks in section II of [65] or Hartry Field's extended discussion in [17].

If, however, part of the worry about Skolem's Paradox stems from the kinds of historical arguments sketched above, then it's easy to see why the parallel between "countable" and "finite" might not seem compelling. In the "finite" case, we might feel perfectly comfortable distinguishing between the ordinary-English sense of "finite" and the sense which is captured (or *mis*captured) by formulas like $\Omega_f(x)$. This is because we don't have any real qualms about the ordinary-English meaning of "finite" to begin with. In the case of "countable," on the other hand, we may well have such qualms. "Countable" is a paradigmatically set-theoretic notion, and the arguments just sketched raise real worries about the determinacy (and even, perhaps, the coherence) of "ordinary-English" set theory.

At this point, then, we have a whole battery of arguments—or, at least, sketches of arguments—against the semantic determinacy of ordinary-English set theory. If any one of these arguments is successful, then there may well be something wrong with the analysis of Skolem's Paradox that I gave in the last two chapters. In particular, since my solution uses the assumption that $\Omega_E(x)$ has a determinate "ordinary-English" semantics, these arguments threaten to turn my solution into mere "naive prattle."

### 3.1.2 A Response

It seems to me that there are three ways that we might respond to the "naive-prattle" objection raised in the last section. The first, and the least satisfactory, would be to simply respond to each of the arguments in the last section individually. We might, for instance, present a counter-history to the one developed on pp. 55–57, a counter-history which makes set theory look more stable and more historically "continuous" than the history previously presented.[20] Similarly, we could try to explain how Klenk has misunderstood the relationship between formal and informal mathematics,[21] and we could argue that the basic assumptions behind Wright's "manifestation argument" are misguided.[22]

[20] So, for instance, we might emphasize the real historical continuities which run from Cantor's set theory to Zermelo's original axiomatization to modern-day ZFC. These continuities might lead us to think that non-standard set theories like Quine's and Wang's are sufficiently out of the mainstream that they shouldn't really count in any historical assessment of that discipline. Also, we could point up some of the intuitive underpinnings of Zermelo-style set theory—e.g., the iterative conception of sets—and we could explain why set theories based on *these* intuitions have never had real problems with the set-theoretic paradoxes. Finally, we could try to explain why set theory should, at this stage in its development, be in the dark about issues involving the continuum hypothesis or the axiom of constructibility. Of course, all this this is only a direction, but it gives an idea as to how an appropriate "counter-history" might begin to go.

[21] We might, for instance, agree that formal axioms are supposed to clarify informal notions, but deny that this clarification has anything to do with the axioms *model-theoretic* properties. Perhaps it's only through *proof theory* that axioms clarify our informal notions—i.e., by making more perspicuous just what these notions really entail.

[22] Here we could argue that the manifestation challenge winds up making *too many* things indeterminate. We could notice, for instance, the ways Dummett uses this challenge to argue against the reality of time (see [14]), or we could look at Putnam's use of the challenge to defend his "cats-and-cherries" argument (see [42], ch. 2). From another angle, we could argue that set-theoretic language actually meets the manifestation challenge (perhaps by developing an account of set-theoretic perception like that in [32] and then discussing the circumstances under which we are "responsive" to particular sets). In either case, our response would undercut the use Wright makes of the manifestation challenge in defending Skolem's Paradox.

Unfortunately, this piecemeal response has two major disadvantages. First, some of the arguments sketched in the last section are genuinely challenging, and it's hard to see just how a convincing response to them would go.[23] More importantly, even if we could respond to *all* of the arguments in the last section, we still wouldn't solve the basic problem. Unless we could ensure that *no further* arguments against the determinacy of ordinary-English set theory were forthcoming, we couldn't be sure that the naive-prattle objection had really been overcome. In short, the piecemeal response to the naive-prattle objection leaves our solution to Skolem's Paradox playing a perpetual game of one-upsmanship against any new objections to ordinary-English set theory which may happen to be developed.

This brings us to a second, and far more satisfactory, response to the naive-prattle objection. To settle matters once and for all, we could simply provide an explicitly worked out semantics for ordinary-English set theory. To do so, we would need to do two things. First, we would need to give an explicit account of what the set-theoretic universe consists of—i.e., an explicit ontology and metaphysics for set theory. Next, we would need to explain *how* ordinary-English expressions like "all sets" and "is a member of" hook up to this universe. That is, we would need to provide an adequate account of the mechanisms whereby the terms and phrases of ordinary English gain semantic significance. By providing all this, we would show fairly definitively that our solution to Skolem's Paradox is something more than mere "naive prattle."

However, although this response would clearly defuse the naive-prattle objection, its very size and complexity make it seem like overkill in the present context. Surely we shouldn't have to accomplish everything suggested in the last paragraph simply to get a solution to Skolem's Paradox! This, then, brings us to a third line of response to the naive-prattle objection. Taking our cue from the shear difficulty of answering the arguments presented in the last section—whether we do this piecemeal or by giving a full semantics for set theory—we might ask whether these arguments really need to be answered in the first place. It seems to me that, in the final analysis, these arguments *don't* need to be answered.

To see why they don't need to be answered, we need to go back and recall the original *point* of Skolem's Paradox. Skolem's Paradox is supposed to highlight an inconsistency (or perhaps an *incoherence*) in our ordinary ways of thinking about set theory. In particular, it's supposed to show that there is a conflict between the naive acceptance of Cantor's theorem and (some instances of) the Löwenheim-Skolem theorems. Since the Löwenheim-Skolem theorems are, presumably, unassailable, this leads to the conclusion that naive talk about "absolutely uncountable sets" is to be avoided.[24]

[23] In particular, although I think it's fairly clear *that* my solution to Skolem's Paradox can be modified to fit into a structuralist framework, I think it's far less obvious *how* these modifications would actually work. For one thing, spelling out the modifications in detail would first require developing a fairly complete formulation of set-theoretic structuralism within which to work. Those familiar with the literature on this topic will see that this is a daunting project (see [23], [24] and [37] for more details).

[24] It is at this point, of course, that formulations of Skolem's Paradox begin to diverge quite wildly. Some take the abandonment of naive set theory as grounds for adopting some kind of formalism. Others take it to show "every set is countable from some perspective" and that we should look for alternate set theories which respect this maxim (e.g., Wang's system). As I argued in the introduction, however, none of these more detailed arguments can get off the ground until *after* we have exposed the initial "conflict" between Cantor's theorem and the Löwenheim-Skolem theorems.

Notice the order of argument here. We are supposed to start with a naive acceptance of Cantor's theorem. (At the very least, we start with an open mind towards this theorem and towards the naive set theory which accompanies it.) We then formulate Skolem's Paradox, discover that there is a problem with our initial naiveté, and therefore abandon our original acceptance of naive—or "ordinary-English"—set theory. In short: Skolem's Paradox is supposed to do the philosophical work here, and the overthrow of ordinary-English set theory is supposed to be the philosophical payoff.

However, once we accept some version of the naive-prattle objection, we reverse this order of argument. Instead of beginning with Skolem's Paradox and using this paradox to show that there's a problem with ordinary-English set theory, we begin with a rejection of ordinary-English set theory and use this rejection to bolster Skolem's Paradox. Diagrammatically, we want to give an argument which looks something like this:

$$\text{Skolem's Paradox} \Longrightarrow \text{ordinary-English set theory is unacceptable.}$$

But, as the argument has actually developed, it looks far more like this:

$$\begin{aligned} \text{Various Arguments} \quad &\Longrightarrow \quad \text{ordinary-English set theory is unacceptable} \\ &\Longrightarrow \quad \text{Skolem's Paradox works} \\ &\Longrightarrow \quad \text{ordinary-English set theory is unacceptable.} \end{aligned}$$

Clearly this latter argument is far less impressive than the first. If we begin with a rejection of ordinary-English set theory, then it's not too surprising that we can, after taking a little detour through Skolem's Paradox, wind up with an argument against the ordinary-English use of "uncountable."

From the standpoint of investigating Skolem's Paradox, then, there are two problems with the naive-prattle objection. First, because the objection starts with a rejection of ordinary-English set theory—starts, that is, with the very thing that Skolem's Paradox is supposed to establish—it renders Skolem's Paradox itself superfluous. In the presence of any of the arguments discussed in the last section, we can argue directly against ordinary-English understandings of "uncountability" without mentioning Skolem's Paradox at all. In effect, then, the naive-prattle objection reduces Skolem's Paradox to mere technical window-dressing.

Second, the naive-prattle objection seems to require that we solve *all* problems with ordinary-English set theory before we can be said to solve *any* of them. Since the solution to any particular problem will, presumably, involve a little bit of set theory (or, at least, using words like "set" and "membership"!), an objector could always respond to the solution by raising *other* problems about the set theory in question. If the ability to raise these *new* problems entailed that our original solution was inadequate, then *no* problem would ever have an adequate solution. Since this conclusion is clearly ludicrous, the naive-prattle objection should be abandoned.[25]

[25] It is as though someone required us to solve both the liar paradox and Descartes' puzzle about the "evil deceiver" before we tackled problems in ethics. After all, when we investigate ethical problems, we might want to use the word "truth" and/or refer to ordinary material objects (e.g., Bill Clinton)!

I propose, therefore, that the proper response to the naive-prattle objection is to sidestep it. As long as there is no explicit conflict between the Löwenheim-Skolem theorems and "ordinary-English" set theory, there is no reason to worry about Skolem's Paradox. Further, in showing that these two (purportedly conflicting) pieces of mathematics are compatible, we should feel free to use all the ordinary-English set theory we want.[26] We do so, not because we think that there are are no other problems with ordinary-English set theory, but simply out of a recognition that these are, in fact, *other* problems. Hence, their solution is not a prerequisite to the use of ordinary set theory in solving Skolem's Paradox.

### 3.1.3 The Paradox of the Three Men

Since this "sidestepping" response to the naive-prattle objection may seem like a bit of a cheat, I'd like to end this section by constructing an analogy that will make the response seem more plausible. The structure of my argument will be simple. I'll begin by presenting, and then solving, a puzzle which bears a close resemblance to Skolem's Paradox. I'll then notice that a variant of the naive-prattle objection can be raised with respect to my solution, but that, in this particular case, the objection is *clearly* misguided. Finally, I'll argue that, by parity of reasoning, the original naive-prattle objection must also be misguided.

My puzzle goes as follows. Three men walk into a bar. They order drinks. At the end of the evening, the bartender presents them with a bill for $30. Each man pays $10, and the three men leave the bar. Once they have left, the bartender realizes that the bill should only have been $25, so he gives the busboy $5 to take to the three men. The busboy, being a dishonest chap, gives each of the men $1 and pockets the remaining $2 for himself. At the end of the day, then, each of the three men has paid $10 and gotten $1 back. Now, $10 - 1 = 9$, and $3 \times 9 = 27$. So, if we add in the $2 in the busboy's pocket, we are left with $29.

However, since we started our story with $30, something has clearly gone wrong. For our purposes, it doesn't really matter what this "something" is. Perhaps it involves problems with our traditional understanding of the arithmetic operators (e.g., addition, subtraction and multiplication); perhaps it involves problems with the identity of the numbers themselves (maybe there is a hidden sorites paradox here); perhaps it involves problems with the way we *refer* to specific numbers (maybe there's an indeterminacy in our use of the numerals "29" and "30"). In any case, the fact that we can start with $30 and wind up with $29 shows that there's *some* problem with traditional "naive" arithmetic. As sophisticated philosophers, therefore, we should eschew such arithmetic in our philosophical deliberations.

This, then, is the paradox of the three men. It's solution is relatively simply. As suggested above, the three men *did* pay $27. This $27 can be split into two parts: $25 which is inside with the bartender and $2 which is in the busboy's pocket.[27] There is also $3 which the men received back from the bartender (via the busboy). As expected, $27 + 3 = 30$. The puzzle turns, therefore, on a simple trick. Instead of adding the $3

[26] In 4.4.4 I present another justification for this attitude towards the use of ordinary set theory in solving Skolem's Paradox. See the discussion of conditionals on pp. 102–103.

[27] The $2 is, after all, money which the three men originally gave to the bartender and which they never got back.

which the men received to the $27 which they payed, the puzzle tries to convince you to add the $2 which is in the busboy's pocket to the $27 and to simply ignore the $3. Since the $2 is *already part* of the $27, this is clearly illegitimate.

Notice, however, that an objection, structurally analogous to the naive-prattle objection, can be made to this solution. After all, the solution makes (naive) reference to all sorts of natural numbers: 2, 3, 25, 27, 29 and 30. It also employs higher-order operations on these numbers: addition, subtraction, multiplication, etc. Further, there are clearly reasons to be skeptical of naive arithmetic. We might, for instance, appeal to the arguments Benacerraf gives in [3] to motivate worries about the determinacy of naive talk about the natural numbers. Similarly, we could appeal to the arguments in [2] to motivate worries about how we could come to *know* about natural numbers (since, we don't, presumably, have causal interactions with natural numbers). Finally, we could appeal to physicalist worries: unless there happen to be infinitely many things in the material universe, it's not clear that there is enough *stuff* to provide a satisfactory ontology for ordinary arithmetic (see [17]). In light of these skeptical worries, then, it might seem as though my solution to the paradox of the three men—relying, as it does, on a good bit of ordinary arithmetic—amounts to just "so much naive prattle."

Clearly, however, something is wrong with this analysis. The paradox of the three men is not a genuine paradox, and the solution given above explains why it is not a genuine paradox. The mere fact that we can raise *other* worries about arithmetic does not show that *this particular* paradox is well-conceived; nor does it show that we cannot use ordinary arithmetic to explain why the paradox isn't well-conceived. In short: in the context of solving a puzzle about arithmetic—i.e., of showing why a particular arithmetical argument does not, in fact, lead to a contradiction—the use of arithmetic is both expected and unproblematic. In this context, therefore, the naive-prattle objection is simply misguided.

The same thing can be said about the original naive-prattle objection. To solve Skolem's Paradox, we need only show that there is no conflict between ordinary-English set theory and the Löwenheim-Skolem theorems. In showing this, we can use all the ordinary set theory we want to. We do not need to give a preliminary defense of this set theory before we begin; nor do we need to answer miscellaneous objections to this set theory. As long as we show that the Löwenheim-Skolem theorems do not, themselves, lead to a contradiction with our set theory, we have done all that is necessary to resolve Skolem's Paradox. In this context as well, therefore, the naive-prattle objection is simply misguided.

## 3.2 Axioms and Mathematical Content

At this point, I turn to a second objection to the solution to Skolem's Paradox which I developed in chapters 1 and 2. The basic structure of this objection is simple. The objection first argues that, for both historical and philosophical reasons, the semantic content of ordinary-English set theory should be *identified* with the content captured by first-order model theory. It then argues that this identification entails that $\Omega_E(x)$ and

$\Omega_M(x)$ share the same semantics. As a result, it concludes that ($\dagger'_M$) winds up being true and that argument (A) winds up being sound.[28]

Clearly, both parts of this objection deserve closer consideration. Without further elaboration, it's not clear why the semantics of $\Omega_E(x)$ should be viewed as having *any* relationship to first-order model theory; further, even if we grant that they have *some* relation to first-order model theory, it's not clear why they should be identified with the semantics of any *particular* $\Omega_M(x)$.[29] Since the first part of this objection is clearly the heart of the matter—and since I don't think this first part is plausible enough to make the second part particularly relevant—I begin by examining the general arguments for relating $\Omega_E(x)$ to (some form of) first-order model theory.

### 3.2.1 Axioms and Implicit Definition

It seems to me that there are two kinds of reasons for thinking that the semantics of $\Omega_E(x)$ are—or, at least, could be—related to model theory. Both involve the roles which axioms play in formalizing and clarifying our mathematical notions. The first involves examining these roles from a historical perspective. It begins by noting that set theorists often resort to axiomatization to show that set theory can avoid various problems or paradoxes. So, for instance, Zermelo's original axiomatization of set theory was a direct response to Konig's "proof" that the set of real numbers cannot be well-ordered: Zermelo believed that *every* set can be well-ordered, so he constructed an axiomatization of set theory in which this result could be proved, but in which Konig's competing result could not be proved.[30] Similarly, when Quine wanted to show that there was nothing problematic about a universal set—i.e., about a set of all sets—he provided an axiomatization of set theory in which the existence of such a set could be proven, but which did not allow the (obvious) derivation of Russell's paradox (see [46]).

These examples highlight the ways in which set theorists use axiomatization as a means for clarifying and making precise the notions with which thy have been working (especially when these notions come under pressure from various sorts of paradox). The historical significance of axiomatization can be further highlighted by noticing the degree to which ZFC has become, at least partially, *constitutive* of the mathematical notion of set. Note, for instance, that anyone who works in a set theory which conflicts with ZFC is taken to have a "non-standard" conception of sets. More tellingly, we often find philosophers arguing about whether various informal conceptions of sets can be adequately squared with our (assumed) commitment to ZFC. In [6] and [64], for example, Boolos and Wang respectively argue that the so-called "iterative conception of sets" provides a good informal justification for the axioms of ZF. In [38], Pollard challenges this argument, claiming that the iterative conception cannot provide a good resolution of the Burali-Forti paradox and,

[28] So, in contrast to the naive-prattle objection, this objection should be understood as a direct *defense* of ($\dagger'_M$).

[29] e.g., why not identify them with $\Omega_1(x)$? Why identify them with $\Omega_M(x)$ rather than $\Omega_{M'}(x)$?

[30] Or, at least, could not be proved in the way that Konig had originally proved it. See [29] for Konig's proof and [68] for Zermelo's response (including his initial axiomatization of set theory). See [35] for an interesting discussion of the whole incident.

hence, must be weaker than ZF. Similarly, Boolos and Menzel have disagreed about whether the iterative conception can really justify the axiom of choice (see [6] and [34]).

For our purposes, the interesting thing about these discussions is the degree to which ZFC is taken for granted. The philosophical question is not whether ZFC can live up to some normative, philosophically-developed conception of sets. Instead, the question is whether we can find a philosophical explication of the set theory we already have—i.e., ZFC. In this context, ZFC serves as a kind of fixed-point for our reflections: we *know* that ZFC is essential to our understanding of sets and we simply want to find out whether a philosophical analysis can deepen this understanding. The fact that ZFC has this fixed-point status provides yet another reason for thinking that axiomatization plays a crucial role in clarifying—and perhaps even *constituting*—our ordinary notion of sets.[31]

These, then, are two historical reasons—or, better, *sociological* reasons—for thinking that axioms are central to our understanding of set theory. From a more philosophical perspective, we should focus on the role axioms play in (at least a certain conception of) mathematical *rigor.* One of the things which distinguished mathematics from other branches of learning is the degree to which the notions of mathematics have been formulated "rigorously." Given a mathematical claim, we can usually say fairly precisely what that claim means: we can break down it's notions into their constituent parts, we can explain (exactly) what kinds of things this claim entails, and, while we may not have a proof of the claim ready-to-hand, we can usually explain pretty well what such a proof would consist in.

This rigor—or precision—can often be attributed to the use of axiomatization in mathematics. When we axiomatize, we force ourselves to get clear about what the basic concepts and principles of a particular discipline are, and we provide a notational framework for defining our less-basic notions. We also, at least if we axiomatize in a system which has a worked-out proof theory, commit ourselves to a specific standard of proof—i.e., we accept precise rules governing what does and does not *count* as "proving something" from our axioms.[32] Finally, we provide an important mechanism for tracking down *errors* in our reasoning: because axiomatization employs a precise notion of proof, it makes it relatively easy to figure out *where* an error has crept into a particular mathematical argument. In all of these ways, then, axiomatization helps to ensure the rigor of pure mathematics.

In the case at hand, therefore, the very fact that set theory is a rigorous mathematical discipline seems to ensure that the content of terms like "set" and "membership" is captured—or at least can be captured—by axioms. Axiomatization is the standard way of ensuring mathematical rigor. Set theory is a rigorous

[31] Nothing here should be taken to imply that there cannot be *more* to our notion of sets than that which is given by ZFC. Clearly, part of the purpose of the discussions mentioned above is to find something more. Also, we may find that, over time, new axioms get added to our standard axiomatization of set theory—e.g., large-cardinal axioms. In either case, however, ZFC remains a *part* of our understanding of sets.

[32] In practice, of course, we very seldom write our proofs out in full detail. Nevertheless, the fact that we *can* write them in such detail—and that we have a standard governing what *counts* as "full detail"—contributes to the rigor of mathematics.

mathematical discipline.[33] Hence, in some manner or another, axioms must capture the real content of set theoretic language.[34]

Before developing this line of argument any further, a brief comment is in order. The arguments I have just sketched are not intended to *contrast* axiomatic set theory with ordinary-English set theory (showing, I suppose, that axiomatic set theory is good while ordinary-English set theory is bad).[35] Instead, they are supposed to show that axiomatic presentations of set theory are at least partially constitutive of the modern, ordinary-English notion of sets. The historical arguments show *that* axiomatic set theory has gained this special status, while the philosophical arguments explain *why* this status is warranted—i.e., because of the imperative to make set theory rigorous.

If these arguments are correct, then, ordinary-English talk about sets serves to *abbreviate* (substantially more cumbersome) talk within a fully axiomatized set theory—i.e., it abbreviates expressions written in a formal language and evaluated against the backdrop of an explicit set of set-theoretic axioms. Because the expressions of ZFC are so messy to write down (see fn. 19), it is more convenient to use phrases like "uncountable" or "is measurable" then to use the (completely formal) expressions that these phrases abbreviate. In saying this, we neither criticize ordinary-English set theory nor imply that it lacks semantic content: we simply explain *how* ordinary-English set theory gains the semantic content that it actually has.[36]

Of course, saying that ordinary-English set theory gains its semantic content through some relationship to axiomatized set theory leaves it unclear just *how* this process is supposed to work. How, for instance, does axiomatized set-theory get *its* semantic content? How *exactly* is this content transmitted to ordinary-English talk about sets. For our purposes, however, there is really only one way to answer such questions. To make Skolem's Paradox work, we are going to have to say that 1.) the semantics of ordinary-English set theory are identified with those of axiomatized set theory and 2.) the semantics of axiomatized set theory are fixed by the model theory of first-order ZFC.

The idea, then, is something like this. The semantics of a phrase like "$x$ is measurable" are fixed by what remains constant when we interpret the formal analog of this phrase across different models of ZFC. More concretely, we start with a specific formal expression corresponding to "$x$ is measurable," and we examine the elements that this expression picks out in various models of ZFC. If all such elements share a common

[33] And one which actually does make extensive use of axioms in the course of its day-to-day practice.

[34] This is not to say that there is no room for intuition in set theory. We need intuition to tell us, for instance, *which* axioms we want to accept. It's just to say that, as set theory has developed into a serious mathematical subject, intuition has gradually given ground to axiomatics. This is an inevitable part of the "maturation" of set theory as a genuine—i.e., rigorous—mathematical discipline.

[35] *This* line of argument was discussed in the last section. See 3.1.1.

[36] A proponent of this position might take the analysis at the beginning of 1.2.1 to illustrate his position. He would first claim that sentences like "$x$ is uncountable" get their semantic content *from* their relation to sentences like Cantor($x$). He would then claim that Cantor(x) gets *its* content from a relation to the formula $\Omega(x)$ and that $\Omega(x)$ gets its content from being embedded within a larger formal system. How, exactly, this last step is supposed to work is somewhat mysterious (and will be discussed shortly); the remainder of the analysis, however, seems fairly clear.

structural property—say, having infinitely many "members"—then this property is part of the *content* of "$x$ is measurable."[37] If some property is not shared by all such elements, then this property is not a genuine part of the content of "$x$ is measurable." Similarly, but more importantly, only those properties which remain constant when we interpret $\Omega(x)$ at different models of set theory count as a genuine part of the content of "$x$ is uncountable."

This idea could be motivated by noting that something very similar goes on in other branches of mathematics. In algebra, for instance, we write down a series of axioms to characterize—or perhaps to "implicitly define"—the notion of "group." Any structure which satisfies these axioms counts as a group and, hence, as a suitable context for interpreting other definitions in group theory. So, for example, we can write down formulas which say—at least against the backdrop of the axioms of group theory—that a particular element is the identity, that a particular element has order five, or that a particular element lives in the center of a group. We can also write down axioms which characterize the notion of an "element with infinite order." In a very real sense, then, these axioms and formulas provide the content of phrases like "is a group," "has order five," or "has infinite order."[38]

There are two things to note about this example. First, the use of axioms to characterize the notion of "group" leads to a kind of "relativity" that is similar to the "relativity" that is often discussed in connection with Skolem's Paradox. "The identity element," for instance, is a phrase which is relative to the particular group about which we are speaking: it is possible for my cat Gandalf to be the identity element of group $G$, to have infinite order in group $H$, and to be completely absent from group $A$. Nor is there an "absolute," ordinary-English sense of "identity element" which will solve this problem—a sense of "identity element" which will determine whether Gandalf "really is" the identity. Instead, the entire content of the phrase "$x$

[37] In this case, if we interpret the notion of "membership" as being relative to the particular model in question—e.g., as discussed in 1.1—then having infinitely many members will indeed be a part of the content of "x is measurable."

[38] Philosophers sometimes write as though the axioms in algebra serve primarily as a basis for deductions. We start by writing down a series of axioms, then we see what conclusions "follow" from these axioms—i.e., which conclusions can be deduced from the axioms in some formal system. This position has been dubbed "if-thenism" or "deductivism." See [48], ch. 3 or [39] for more on this position.

However, this position is fairly clearly false, even as applied to axiomatic disciplines like group theory. While the axioms of group theory suffice to prove fairly trivial results about groups—e.g., every group has a unique identity, every element has a unique inverse—they don't take us through (even) the first week of a typical undergraduate course in the subject. They will not, for instance, let us prove theorems which *compare* various groups—e.g., which let us show that one group is a subgroup of another, that two groups are homomorphic to each other, or that one group is the quotient of two other groups. Indeed, theorems of this kind can't even be *formulated* in the language in which the axioms of group theory are usually written. As a result, the axioms of group theory play only a minor role in the actual *proof* of theorems like the first isomorphism theorem—theorems which involve stepping outside the groups we are interested in to discuss the structural relationships *between* these groups.

In practice, then, the axioms of group theory serve almost purely to *pick out* the structures about which group theorists intend to talk—or, in other jargon, to "implicitly define" the concept of group. Once these structures have been isolated, group theorists step back to talk about these structures in a relatively non-axiomatic manner. This fits far better with the account of axioms sketched in the main text than with the account suggested by "if-thenism."

is the identity element" is bound up with the way an associated bit of formalism—$\forall y\, [x \cdot y = y \cdot x = y]$—gets interpreted in different models of group theory. Since these interpretations vary, the phrase itself is unavoidably "relative."[39]

Second, none of this is peculiar to group theory. Field theorists use axioms to characterize the notion "field of characteristic zero," and ring theorists use axioms to characterize the notion "divisible ring." Outside of algebra, topologists use axioms to characterize the notion of a "topological space," and geometers use the Eilenberg-Steenrod axioms to characterize the notion of a "homology theory." In each of these cases, new formulas can be used to define additional concepts against the backdrop of the structures picked out by the initial axioms. Hence, far from being some kind of oddity, the use of axioms to characterize—or to "implicitly define"—mathematical notions is quite widespread.

At this point, then, we have a rough outline of an argument which serves to bolster Skolem's Paradox. We begin by arguing that ordinary-English set theory gets its semantic content through its association with axiomatized set theory. We then argue that axiomatized set theory gets its semantic content from its model theory (spelling this claim out by appealing to an analogy with cases from, e.g., algebra). At the end of the day, we conclude that the content of ordinary-English talk about "sets" and "membership" is fixed, in its entirety, by the model theory of first-order ZFC. Once this position is on the table, we have a clear rationale for taking puzzles like Skolem's Paradox (far) more seriously than we did before.

To tie this argument to the material from chapters 1 and 2, we should notice two things. First, the argument provides a justification for being skeptical about my assumption that the "ordinary-English" semantics of $\Omega_E(x)$ can be blithely distinguished from the formal semantics of sentences like $\Omega_1(x)$ and $\Omega_M(x)$. If the present argument is correct, after all, then something like these latter semantics will have to *be* the semantics of $\Omega_E(x)$. Second, the present argument lays the groundwork for further arguments which would, at least in certain contexts, identify the semantics of $\Omega_E(x)$ with those of $\Omega_M(x)$. If these more detailed arguments go through, then Skolem's Paradox will also go though (as shown in 1.2.2).

### 3.2.2 Why Model Theory?

There are two concerns that I want to raise about the line of argument just sketched.[40] The first involves the transition between the view that axioms are constitutive of our ordinary notion of sets and the view

[39] The same point holds for other phrases. It is entirely possible to build a sequence $g \in G \subset G' \subset H$ such that 1.) $g$ is in the center of $G$ but not in the center of $G'$ or $H$ and 2.) $G$ is a maximal subgroup of $G'$ but not of $H$. This shows that notions like "center" and "maximal subgroup" are also relative to the particular group under discussion.

[40] There are, of course, more than two concerns that *could* be raised here. We could, for instance, raise concerns about the initial argument for viewing axioms as *constitutive* of our ordinary notion of sets. We could also ask for a lot more detail about the exact ways in which first-order model theory "captures" the content of phrases like "$x$ is uncountable." Finally, and as noted on page 63, we could raise concerns about whether the mere argument that set theory has a *model-theoretic* semantics can really be parlayed into argument for *identifying* the semantics of $\Omega_E(x)$ with those of any specific $\Omega_N(x)$. For the present, however, I think it is best to sidestep these concerns and focus solely on the issues mentioned above.

that it's the *model theory* of these axioms that does the actual constitutive work. In principle, the mere assumption that axioms play some role in fixing the content of our notion of set, combined with the fact that some axioms fix content though model theory, is not enough to ensure that the axioms of set theory fix content through model theory. Further, there are at least two reasons for being skeptical of the idea the axioms of set theory fix content in this manner.

First, if we look at the original arguments for taking axioms to play a constitutive role in fixing the notion of "set," we find that these arguments have more to do with the proof-theoretic properties of axioms than with their model-theoretic properties. It was by looking at what ZFC can prove—or, more precisely, what kinds of proofs *don't* work in ZFC—that mathematicians determined that ZFC doesn't generate Russell's Paradox, Konig's Paradox or the Burali-Forti Paradox. Similarly, it was the proof-theoretic properties of axioms which helped to explain how axioms contribute to mathematical rigor. In no case did anything model-theoretic enter into these initial arguments for connecting axioms to the content of set-theoretic language.

Given this, why shouldn't we think that the proof-theoretic properties of axioms do most (or even all) of the work in explaining how axioms contribute to our understanding of sets? We might, for instance, think that the axioms of set theory serve to state our basic intuitions about sets. To determine what these intuitions amount to—i.e., what the *content* of these axioms really is—we look to see what the axioms of set theory are able to prove (rather than looking at what remains stable across all models of set theory). So, the fact that the axiom of extensionality is among our axioms ensures that extensionality is part of the content of our notion of set (even though the same *object* may get different "members" as we move from one model of set theory to another). Likewise, the fact that ZFC proves Cantor's theorem is enough to ensure that our conception of sets is rich enough to include uncountable sets. The fact that the model theory of ZFC fails to "capture" this notion of uncountability—i.e., fails in the ways exposed by Skolem's Paradox—is simply beside the point.

Clearly, this picture would need a lot of elaboration before it would constitute an adequate account of the way axioms contribute to the content of ordinary-English set theory. But even this sketch is enough to highlight my basic point: the (mere) assumption that axioms play some role in fixing the content of our notion of "set" doesn't show that there is a specifically *model-theoretic* process by which this content-fixing takes place. As long as there are other processes on the table—and, in particular, processes that are better tied to our initial arguments for looking at axioms in the first place—we should be quite skeptical of the claim that the content of ordinary-English set theory is fixed by the model theory of first-order ZFC. Hence, we should also be skeptical of the thought that reflection on the role of axioms in mathematics will provide support for Skolem's Paradox.

Second, there are obvious dissimilarities between the ways mathematicians think about the axioms of set theory and the ways they think about the axioms of subjects like group theory (i.e., subjects where it is *clear* that axioms provide "implicit definitions" of mathematical concepts). For one thing, mathematicians *do* tend to talk as though there is an absolute understanding of "set" which determines, once and for all, which objects count as sets. They do not engage in similar talk about “group elements.” So, for instance, few mathematicians would be willing to say that my cat Gandalf is a set, still fewer that he is some particular set like $\aleph_{17}$.[41] In contrast, it’s absolutely uncontroversial that Gandalf is an element of many groups and that he is the identity element in at least some of these groups. This shows that there is a real difference between the ways mathematicians think about set theory and the ways they think about group theory.

For another thing, mathematicians tend to treat the axioms of set theory as being less *fixed* than those of group theory or topology. In set theory, mathematicians sometimes raise the question as to whether the ZFC axioms are correct—i.e., they talk as though there is an intuitive notion of set against which the ZFC axioms might be checked and found wanting. In group theory and topology, it simply makes no sense to talk about “intuitive notions” which could diverge from the notion specified by the relevant axioms.[42] In a similar vein, set theorists sometimes debate whether we should add new axioms to the standard axioms of set theory—e.g., large cardinal axioms, or axioms like V=L, or even just axioms like Con(ZFC). In contrast, no one would dream of making additions to the axioms of group theory or topology.

I conclude, therefore, we have no reasons for thinking that the content of ordinary-English set theory is fixed by the model theory of ZFC. Even if we grant that ZFC plays an important role in fixing the significance of ordinary talk about sets, we have no reasons for thinking that this works by way of *model-theory.* For one thing, alternate accounts of how ZFC could fix such significance are clearly available, and some of these accounts fit better with our reasons for thinking that axioms play *any* role in fixing content. For another, there are obvious dissimilarities between the ways axioms are treated in set theory and the ways they are treated in other disciplines which involve some form of axiomatic “implicit definition.” Given this, we should be quite skeptical of the move from the view that axioms are constitutive of our ordinary notion of sets to the view that it’s the *model theory* of these axioms that does the actual constitutive work.

### 3.2.3 The Supermodel

In this section, I turn to a second concern about the argument sketched in 3.2.1. Even if we assume that the axioms of set theory *do* function to provide “implicit definitions” of set theoretic concepts, this still won’t be enough to ground the position sketched on pp. 65–67. To see why, we need to look a little closer at the ways axioms can be used to “pick out” mathematical structures and/or concepts. To help this examination along, I begin by describing a particular (and rather peculiar) model and looking at this model’s relationship to several sets of axioms.

The model I want to discuss—which I shall call the “supermodel”—is relatively straightforward. It’s domain consists of a single object (Tim), and its “membership” relation is defined so as to be as large as

[41] Presumably, they *would* say something like this—at least in some contexts—if they thought that all it took to *be* a set was to be an element of some model of ZFC.

[42] As noted earlier (p. 63), some philosophers take a similar line about ZFC. But, in the case of ZFC, it is at least an *open question* whether this is the right line to take. In the case of group theory or topology, it is the *only* line which makes any real sense.

possible (given this domain). More formally,

- $S = \{\text{Tim}\}$
- $\in_S = S \times S$
- $\mathbb{S} = \langle S, \in_S \rangle$

Since the supermodel is supposed to be a model for the language of set theory, this is all we need to say to give a complete specification of that model.

Now, the supermodel comes with it's own satisfaction relation—which I shall call "supersatisfaction"—and it is this (non-standard) satisfaction relation which makes the supermodel unusual. The definition of supersatisfaction is exactly like that of ordinary (first-order) satisfaction, except that negation is stipulated to be a redundant operator. More formally, the recursion clause for ¬ reads: for any formula $\phi$ and any assignment of variables $\nu$,

$$\mathbb{S}, \nu \models_s \neg\phi \iff_{\text{Def}} \mathbb{S}, \nu \models_s \phi. \tag{3.1}$$

Thus, so long a formula does not use negations, supersatisfaction treats that formula just the way ordinary satisfaction treats it. Once a formula does use negations, however, supersatisfaction treats it differently.

Before discussing the supermodel's philosophical significance, I need to make three technical points about the model and its associated "satisfaction" relation. First, and most important, the supermodel supersatisfies everything. That is, for any formula $\phi$ in the language of set theory and any assignment of variables $\nu$,

$$\mathbb{S}, \nu \models_s \phi.^{43} \tag{3.2}$$

Hence, if we redefine the notion "$M$ is a model of $T$" in terms of supersatisfaction (rather than ordinary satisfaction), then $\mathbb{S}$ will be a "model" for *every* theory $T$.

Second, the fact that the supermodel supersatisfies everything—i.e., fact 3.2 above—can be preserved through various kinds of tampering. So, for instance, we can change the size of $\mathbb{S}$ or even the underlying language of $\mathbb{S}$ while maintaining the fact that $\mathbb{S}$ supersatisfies all formulas. To illustrate this point, and to lay the groundwork for several later points, I provide four brief examples.

First, suppose we want to expand the supermodel by adding new predicates, constants, and functions to its language. As all constants must be mapped to Tim, and as functions are trivial in a one-element model, the only potential choices involve our new relation symbols. And, if we expand the supermodel so as to make

[43] The argument for this claim involves a straightforward induction on the structure of $\phi$. The fact that $S$ contains only one element, along with the fact that $\in_S$ is maximal, ensures that $\mathbb{S}$ supersatisfies all atomic formulas. Given this, the passage through → is trivial (since → maps $(T, T)$ to $T$), and the passage through quantifiers is equally trivial since $\mathbb{S}$ contains only one element (so, $\mathbb{S} \models_s \phi[\text{Tim}] \iff \mathbb{S} \models_s \exists x\, \phi(x) \iff \mathbb{S} \models_s \forall x\, \phi(x)$). Finally, the fact that negation is redundant allows us to move from $\mathbb{S} \models_s \phi$ to $\mathbb{S} \models_s \neg\phi$.

Note that this argument continues to work even if we take ∨, ∧, and ↔ as logical primitives (since each of them also maps $(T, T)$ to $T$). For the remainder of this section, therefore, I will go ahead and treat ∨, ∧, and ↔ as though they were logical primitives (instead of using them to abbreviate more complicated expressions involving only ¬ and →).

*all* relations maximal—i.e., if we make the expanded supermodel interpret *all* predicates by $S$ and *all* n-ary relations by $S^n$—then we can preserve the fact that $\mathbb{S} \models_s \phi$ for every $\phi$ in our (expanded) language.[44]

Second, suppose we want to increase the size of the supermodel by replacing the one-element set {Tim} with some arbitrary set $S$. In this case, our original definition of $\models_s$ does not ensure that 3.2 continues to hold: if there exist $s_1 \neq s_2 \in S$, for instance, then $\mathbb{S} \not\models_s \forall x \forall y(x = y)$. However, if we simply redefine $\models_s$ so as to make equality trivial (i.e., define $\models_s$ so that $\mathbb{M} \models_s m_1 = m_2$ for *any* model $M$ and *any* $m_1, m_2 \in M$), then our new version of $\mathbb{S}$ will, once again, supersatisfy all formulas.[45]

Third, suppose we want to move beyond first-order notation to consider formulas with a more complicated syntax. Clearly, fact 3.2 will not be changed by moving from first-order logic to any logic of the form $\mathcal{L}_{\kappa,\lambda}$ (since quantifiers, conjunctions, and disjunctions are trivial in the supermodel). More significantly, there is a straightforward way of extending 3.2 to cover formulas of second (and higher) order logic. We begin by extending our definition of $\mathbb{S}$ so that every nth-order, m-ary relation is interpreted by $[\mathcal{P}^{n-1}(S)]^m$. We then extend our definition of $\models_s$ so that, for any sequence of nth-level objects $\langle S_1, \ldots, S_m \rangle$, any $n + 1$st-level m-ary relation variable $R$, and any assignment of variables $\nu$,

$$\mathbb{S}, \nu \models_s R[S_1, \ldots, S_m]. \tag{3.3}$$

Together, these two modifications ensure that all atomic formulas (even higher-order atomic formulas) wind up being supersatisfied by $\mathbb{S}$. This supersatisfaction then extends to more complicated formulas just as it did in the last two examples (cf. footnotes 45 and 44).

Fourth, suppose we want to accomplish all three of the tasks from the previous examples *at the same time*. Then, if we simply combine the modifications to $\mathbb{S}$ and $\models_s$ that were discussed in the previous examples, we obtain a variant of the supermodel and supersatisfaction relation such that: 1.) the supermodel uses an arbitrary set $S$ for its domain, 2.) the supermodel's interpretation function "knows about" an arbitrary collection of primitive constant, relation and function symbols (including n-th order relation symbols for any/every finite $n$), and 3.) the supersatisfaction relation deals with arbitrary sentences in the language $\mathcal{L}^{\omega}_{\infty,\infty}$.[46] Further, it is easy to check that the resulting supermodel still supersatisfies (in, of course, a revised sense of "supersatisfies") every sentence in $\mathcal{L}^{\omega}_{\infty,\infty}$. So, fact 3.2 continues to hold.[47]

[44] The argument for this claim is essentially identical to that for 3.2. The fact that all relations are maximal ensures that all atomic formulas come out true (equalities being trivially true in one-element models); from here, we just follow the inductive argument of footnote 43.

[45] As usual, the argument for this claim proceeds by induction on $\phi$. The fact that $\in_S = S \times S$ ensures that $\mathbb{S} \models_s s_1 \in s_2$ for all $s_1, s_2 \in S$; similarly, equalities are taken care of by our new definition of $\models_s$; so the claim holds for all atomic formulas. Given this, propositional connectives work as in footnote 43, and, since this whole argument works for *any* assignment of variables, quantifiers don't change anything.

[46] i.e., sentences which involve (arbitrarily large) infinitary conjunctions, disjunctions and quantificational blocks and which include nth-order quantifiers for any (or, indeed, for all) finite $n$.

[47] The technical details of this construction go as follows. We begin by letting $S$ be an arbitrary non-empty set and letting $\hat{s}$ be an arbitrary element of $S$. We then define an interpretation function such that 1.) all constants are mapped to $\hat{s}$, 2.) all

To recap this second point, then, although the simplicity of the original supermodel makes that model more accessible than the ones we have just been considering, this simplicity plays no *essential* role in proving 3.2. If we employ enough tricks—i.e., if we are creative enough in modifying either $\mathbb{S}$ or $\models_s$ as the need arises—then we can preserve 3.2 through a whole variety of changes in the underlying structure of $\mathbb{S}$. Hence, although I will now revert to using the terms "supermodel" and "supersatisfaction" to refer only to the original model and relation defined on page 70, the flexibility of the supermodel construction should be kept in mind throughout the remainder of this discussion.

This brings me to a final technical point, a point concerning the supermodel's problematic relationship to the principle of bivalence. Clearly, part of the reason $\mathbb{S}$ supersatisfies every sentence is because the supersatisfaction relation fails to conform to bivalence (this, in turn, is a result of the relation's failure to "correctly" interpret the symbol "$\neg$"). It is important, though, that we be clear about just *how* this failure of bivalence really works.

First, note that the supermodel *does* supersatisfy each of the formal sentences which may, at first blush, seem to characterize bivalence. That is, for every sentence $\phi$ in the language of set theory,

$$\mathbb{S} \models_s \phi \vee \neg\phi. \tag{3.4}$$

Of course, this is true simply because the supermodel supersatisfies everything. Hence, to the extent that worries about bivalence are motivated by the thought that $\mathbb{S}$ should *fail* to supersatisfy some things, 3.4 doesn't help very much. In particular, it doesn't help to ensure that, for any sentence $\phi$, $\mathbb{S}$ supersatisfies either $\phi$ or $\neg\phi$ *but not both.*

Although this worry is undoubtedly correct, we should note that the game played with the original version of bivalence can be played in cascades. The formal sentence which corresponds to the demand that $\mathbb{S}$ supersatisfy *only one* of $\phi$ and $\neg\phi$ is, itself, supersatisfied by $\mathbb{S}$. That is, for any sentence $\phi$ in the language of set theory,

$$\mathbb{S} \models_s \neg(\phi \wedge \neg\phi) \tag{3.5}$$

and, indeed,

$$\mathbb{S} \models_s (\phi \leftrightarrow \neg\neg\phi). \tag{3.6}$$

Hence, none of the formal sentences which may initially seem to characterize bivalence can be used to *determine* whether or not $\mathbb{S}$ really conforms to that principle. Because, $\mathbb{S}$ supersatisfies *every* sentence, it clearly supersatisfies these particular sentences. Just as clearly, this fact entails nothing about the supermodel's actual adherence to the principle of bivalence.

n-ary function symbols are mapped to the constant function $f_n : S^n \to \{\hat{s}\}$, and 3.) all nth-order, m-ary relation symbols are mapped to $[\mathcal{P}^{n-1}(S)]^m$. Finally, the relation $\models_s$ gets the same definition as ordinary satisfaction (i.e., ordinary satisfaction for languages of the form $\mathcal{L}^{\omega}_{\infty,\infty}$), modulo the modifications discussed in previous examples: interpreting $\neg$ according to 3.1, interpreting "=" as a maximal relation, and interpreting the predication of higher-order variables according to 3.3.

Given these definitions, it is easy to check that 3.2 holds for all atomic formulas. From there, the induction proceeds as usual to give 3.2 for *all* formulas.

To summarize: the supermodel's failure to satisfy bivalence is not the result of its failure to satisfy—or, rather, to supersatisfy—some particular sentence or sentences in the language of set theory. Instead, it is the result of the supermodel's failure to respect a certain *pattern* in the way it supersatisfies these sentences. In particular, we would like the supermodel and supersatisfaction relation to respect the pattern

$$\mathbb{S} \models_s \phi \Longleftrightarrow \mathbb{S} \not\models_s \neg\phi \tag{3.7}$$

for every sentence $\phi$ in our language. But, as the examples above illustrate, respect for this pattern cannot be reduced to the requirement that $\mathbb{S}$ supersatisfy some particular sentence (or even *collection* of sentences) The supermodel *does* supersatisfy all such sentences and collections, but it manifestly *fails* to respect the pattern at issue.

These, then, are my three technical points concerning the supermodel. 1.) the supermodel supersatisfies all formulas. 2.) this fact has little to do with the supermodel's having a one-element domain or having a language which contains only one symbol (although these facts make the proof of 3.2 somewhat more perspicuous). 3.) the fact that the supermodel and supersatisfaction relation fail to satisfy basic logical principles cannot be cashed out in terms of particular sentences (or collections of sentences) which the supermodel fails to supersatisfy. With these three points under our belt, we can turn to examine some of the philosophical implications of the supermodel/supersatisfaction construction.

### 3.2.4 Philosophical Implications

There are, I think, four lessons to be learned from our examination of the supermodel. First, a model can only satisfy an axiom—or a collection of axioms—against the backdrop of a particular understanding of the "satisfaction" relation. In principle, there are many different ways to understand this relation: as first-order satisfaction, second-order satisfaction, intuitionistic satisfaction, supersatisfaction, etc.[48] These understandings give rise to different answers to questions of the form "does this model satisfy that sentence." The supermodel, for instance, supersatisfies many sentences which it doesn't "satisfy" on any ordinary understanding of "satisfaction."[49] Hence, it is only when we have a specific background semantics for our language—i.e., a specific understanding of what counts as a "model" for our language and a specific account of how these models "satisfy" particular sentences—that it makes sense to talk about a model "satisfying" an axiom or a collection of axioms.

Second, the background semantics for a language are themselves fixed from *outside* that language. To define a specific "satisfaction" relation we need to step back, both from the language in question and from

[48] For another "oddball" satisfaction relation, see section 4.4.2. For non-standard satisfaction relations with more in the way of philosophical and mathematical motivation, see [36], [57] or chapter 2 of [23].

[49] A similar example involves so-called "Henken models" of second-order logic. On a second-order understanding of "satisfaction" these structures don't even count as "models" for second-order formulas; hence, they don't satisfy *any* second-order sentences. On an alternate (first-order, many-sorted) understanding of "satisfaction," they *do* count as "models," and they "satisfy" all kinds of second-order sentences.

the intended models for that language, to discuss the *relationship* between language and models. So, for instance, we need to write down the recursion clauses which determine how our "satisfaction" relation understands symbols like "¬": does it understand "¬" in the manner of ordinary satisfaction, in the manner of supersatisfaction, or in some other manner altogether? These clauses determine whether our satisfaction relation winds up respecting patterns like that in 3.7. It is only from this *external* standpoint, therefore, that we can define the crucial notions—i.e., of models and satisfaction—which make sense of talk about a model "satisfying" a particular axiom.

As a corollary to this point, the background semantics for a set of axioms cannot be specified by means of *further* axioms. We cannot explain why the supermodel constitutes a "bad" interpretation for the axioms of set theory simply by finding other axioms—whether in the language of set theory or in some expanded language—which the supermodel fails to "satisfy."[50] This was the moral of my discussion of the principle of bivalence and of the stability of the supermodel construction under expansions of our language. Instead, we have to step outside our axioms to notice an unfortunate pattern in the way the supermodel relates to negation (i.e., that it regularly supersatisfies both $\phi$ and $\neg\phi$). Nor is this point specific to the supermodel. For any set of axioms, there will be *many* (non-equivalent) ways to interpret symbols like "¬," "→" and "∃" which make the relevant axioms come out "true." Hence, any set of axioms will be "compatible" with a wide variety of background semantics for the so-called "logical" terms.

Third, and as another corollary, it makes no sense to talk in any *generic* sense about axioms "picking out" models and/or "implicitly defining" concepts. If we are willing to deploy enough tricks—i.e., tricks of the kinds discussed on pp. 70–72—then we can make essentially any structure "satisfy" essentially any axioms. So, if we want our axioms to pick out certain models with respect to *every* conception of satisfaction—i.e., to pick out exactly the same models no matter what conception of satisfaction we happen to be using—then *no* axioms will "pick out" *any* models. On the other hand, if we merely want our axioms to pick out models with respect to *some* conception of satisfaction—i.e., if we only demand that there be *a* conception of satisfaction against which our axioms pick out the models in question—then *every* set of axioms will "pick out" *every* model. Without a fixed background semantics for interpreting our axioms, therefore, talk about these axioms "picking out" models or "implicitly defining" concepts is pretty-much meaningless.[51]

This third lesson is important and deserves to be emphasized. Philosophers sometimes talk as though axioms themselves should be able to determine their own semantics. That is, they talk as though, if we could only write down *enough* axioms using a *rich enough* syntax, we could force those axioms to take on a certain meaning.[52] But this is clearly false. The mere act of writing down axioms does *nothing* towards

[50] Since the supermodel will, after all, satisfy *all* further axioms.

[51] Of course, if we *do* have a fixed semantics for some of the symbols in our language—and, at least, a proto-semantics for the other symbols (see 1.2.1)—then it makes sense to ask whether certain collections of axioms "pick out" models and/or "implicitly define" concepts. But, this is very different from the idea that *axioms alone* can do this kind of work.

[52] See, for instance, [28], [43] and [65]. In each case, the authors talk as though the semantics of mathematical language should be *totally* fixed by axioms. If the meaning of these axioms is unclear, then further axioms should be used to "fix" the problem.

fixing the interpretation of those axioms; as we saw in our discussion of the supermodel, it can't even fix the interpretation of symbols like "¬." Hence, the mere act of writing down axioms—no matter *how many* we happen to write or *how much* semantic vocabulary we happen to include—does nothing towards "picking out" a model for these axioms or "implicitly defining" a mathematical concept.[53]

Fourth, this all shows just how hard it will be for the proponent of the line sketched in 3.2.1 to make a convincing defense of Skolem's Paradox. In 3.2.2 we saw that the mere assumption that axioms play *some* role in fixing the content of set-theoretic language does not imply that these axioms' *model theory* does the actual work. Now we see that even the assumption that model theory *does* do the real work won't be enough. Unless we have reason to think that the content of set-theoretic language is captured by the specifically *first-order* model theory of set-theoretic axioms, the argument of 3.2.1 will fall apart.

Of course, the fact that Skolem's Paradox rests on first-order model theory is in no way surprising. This fact was emphasized as far back as section 1.1. But, I hope the discussion of the last few pages has made several things more clear. First, there is no sense in which the axioms of set theory themselves *force* us to understand them via their first-order model theory. Second, the mere fact that the axioms of set theory *can be given* a first-order model theory doesn't show that the semantics of set theory are *no more* determined than this first-order model theory would imply; after all, the mere existence of supersatisfaction doesn't show that the supermodel provides a legitimate "interpretation" of set theory. Third, the decision to treat first-order model theory as canonical for the purposes of fixing content is just that, a decision. There are lots of other—and lots of more attractive—options on the table. Hence, the proponent of first-order model theory needs to give an explicit argument for taking such model theory as the "right" background for fixing the content of set-theoretic language.

What are the prospects for such an argument? As far as I can see, very poor. There are at least three things which tell against the success of such an argument. First, the claim that the meaning of terms like "set" and "membership" is fixed by the first-order model theory of ZFC is *prima facie* implausible. It amounts, after all, to the claim that such meanings can be fixed purely on the basis of a fixed meaning for the symbols $\neg$ and $\rightarrow$ and a proto-semantics which determines that $\in$ is a binary relation and that $\exists$ ranges over *some* domain. But there is simply no reason to think that understanding the meaning of "not" and "if. . . then" (along with some grammatical categories for "membership" and "there exists") *should* be enough to explain the general notion of sets.

Nor do things get better when we turn to, e.g., the transitive submodel version of Skolem's Paradox or the version based on elementary submodels of some $V_\kappa$ (see 2.1 and 2.2). To be sure, these arguments allow us to start with a fixed meaning for "membership." But, as we saw on pages 35–37, it is quite normal for differences in the interpretation of "there exists" to lead to wildly different interpretations of concepts defined by means this phrase. Hence, we have no reason to expect that fixing the meaning of "not," "all,"

[53] The "nothing" here is meant fairly literally. I don't just mean that there is a lot *more* to be done; I mean that the act of writing axioms down doesn't really fix *anything* about what these axioms might ultimately be taken to mean.

"if. . . then," and "membership" (while keeping the phrase "these exists a set" unfixed) will allow us to define notions like "all sets" or "is uncountable" with any degree of precision. In short, then, we have no *prima facie* reasons for thinking that the first-order model theory of ZFC is even a good candidate for fixing the significance of set-theoretic language.

Second, we cannot bolster the case for first-order model theory by examining the ways mathematicians use axioms in other branches of mathematics. In some cases, of course, mathematicians *do* use first-order axioms—e.g., in group theory or field theory. But in many cases, their axioms are *not* first-order. So, for instance, neither the axioms used to define topological spaces nor the axioms used to define linear continuums are first-order.[54] Nor are the Eilenberg-Steenrod axioms in algebraic topology. Thus, no examination of *general* mathematical practice can support a bias towards first-order axioms.

To reinforce this point, note that mathematicians have very good reasons for eschewing first-order axioms (at least in certain circumstances). Many *trivial* concepts of ordinary mathematics just can't be "captured" in first-order model theory. Most famously, and as we noted near the end of 3.2.1, first-order model theory cannot capture the notion of "finitude." As a result, it cannot capture the notion of the *closure* of a set under algebraic operations, nor can it capture recursively defined properties of (or relations on) the natural numbers. Since all of these concepts are *basic* to modern mathematics, mathematicians have good reason to refuse to limit themselves to first-order axiomatization.

Third, even an examination of modern *set theory* won't really support the first-order position we are examining. To be sure, set theorists do make extensive use of first-order ZFC in their research. But set theorists also think about non-first-order axioms. Zermelo's original axiomatization of set theory was second-order, and many of his contemporaries also made use of second-order axiomatizations.[55] Further, the study of second-order ZFC continues to be an active area of mathematical investigation, especially in connection with large cardinal axioms (see [13], [23], and [33] for interesting surveys of this literature).

More importantly, the only reason for *believing* some of the axioms of first-order ZFC is because we already believe certain (related) second-order axioms. In particular, first-order ZFC contains two "axiom schemes" which serve to capture the principles of separation and replacement. In each case, an principle which has a relatively natural second-order formulation gets replaced with an infinite list of (substantially

[54] In the case of topology, we need to go second-order to ensure that *arbitrary* unions of open sets wind up being open. In the case of linear continuums, we need second-order axioms to formulate the right notion of completeness.

[55] For Zermelo's axiomatization, see [68]. For evidence that Zermelo intended this axiomatization to be second-order, see [67]. For other works from the early part of the twentieth century, see [9] or [55].

more complicated) first-order axioms.[56] As far as I can see, however, the only reason to accept these first-order axioms is because we already accept the (far more natural) second-order axioms from which they follow. To the extent that this is right, set theorists' mere use of first-order ZFC doesn't show very much about the actual content of their notion of set.[57]

To conclude, therefore, there is very little reason to accept the claim that the content of set-theoretic language is captured by the first-order model theory of ZFC (even if we *do* accept that this content is captured by *some* form of model theory). The claim is *prima facie* implausible, it gets no support from the general use of axioms in mathematics, and it doesn't even fit very well with the use of axioms in *set theory.* Since it does, on the other hand, tend to lead us into problems like Skolem's Paradox, the claim should simply be rejected.

## 3.3 Conclusion

In this chapter I have looked at two objections to the analysis of Skolem's Paradox presented in chapters 1 and 2. The first was based on a suspicion that my analysis of Skolem's Paradox relied overly-much on a naive acceptance of ordinary-English set theory. It bolstered this suspicion with several arguments against ordinary-English set theory. In 3.1.2, however, we saw that this kind of argument is out of place in the context of solving Skolem's Paradox. Because Skolem's Paradox purports to uncover a problem with (and

[56] In the case of separation, we start with the second-order axiom

$$\forall P \forall x \exists y \forall z [z \in y \longleftrightarrow z \in x \wedge P(z)].$$

To make this axiom first-order, we replace it with a whole series of axioms, one for every formula $\phi(\bar{x}, y)$ in our language. These axioms all have the form:

$$\forall \bar{x} \forall x \exists y \forall z [z \in y \longleftrightarrow z \in x \wedge \phi(\bar{x}, z)].$$

In effect, then, we use an explicit list of formulas to replace the "$\forall P$" in our original axiom, and we use a new axiom to take care of each one of these formulas. A similar strategy allows us to replace the "$\forall R$" in the natural second-order formulation of replacement:

$$\forall R \forall x [\forall y (y \in x \rightarrow \exists ! z R(y, z)) \longrightarrow \exists x' (\forall y (y \in x' \leftrightarrow \exists z (z \in x \wedge R(z, y))))].$$

Instances of the resulting schema look like this:

$$\forall \bar{x} \forall x [\forall y (y \in x \rightarrow \exists ! z \phi(\bar{x}, x, z)) \longrightarrow \exists x' (\forall y (y \in x' \leftrightarrow \exists z (z \in x \wedge \phi(\bar{x}, y))))].$$

Clearly, however, the schemes which result from this kind of substitution are nowhere near as natural and perspicuous as the second-order axioms for which they are substituted.

[57] In practice, I think set theorists like to use first-order ZFC because its proof theory is so perspicuous. In particular, the back and forth between proof theory and model theory guaranteed by the completeness theorem, along with some powerful methods for manipulating the models of first-order ZFC (e.g., forcing), make first-order ZFC a nice place to prove things. Since there are no similar techniques for second-order ZFC, set theorists tend to *work* in the first-order realm.

On my view, then, second-order axiomatization captures the real intuitions lying behind ZFC. Since second-order ZFC entails first-order ZFC, anything proved in first-order ZFC follows from second-order ZFC as well. So, we can use first-order ZFC for the purposes of doing proofs—for the reasons described above—while maintaining that second-order ZFC captures our real commitments.

perhaps even a contradiction in) ordinary-English set theory, it is legitimate to use ordinary-English set theory in solving that paradox. Even if there are other problems with ordinary-English set theory, these problems are topics for another time; the mere existence of other problems does not impune my solution to this one.

The second objection was based on a worry that I had overlooked something about the role which axioms play in fixing the content of our notion of set. Here, I argued for three claims. First, even if axioms *do* fix the content of our notion of set, there is no reason to think that they do this by means of their model theory (and good reasons to think that they *don't* do it this way). Second, even if axioms *do* fix the content of set theory through their model theory, there is no reason to think that it's their *first-order* model theory which does the trick. Third, when these first two points are set alongside the fact that first-order, model-theoretic conceptions of content tend to lead straight into Skolem's Paradox, we obtain a fairly decisive argument for rejecting such conceptions.

At the end of the day, therefore, the analysis of Skolem's Paradox presented in chapters 1 and 2 stands. There is no reason to think that reflection on *other* puzzles in the philosophy of set theory will dislodge this analysis. Nor is there reason to think that reflection on the role that axioms play in mathematics will create any difficulties. Therefore, I take my solution to Skolem's Paradox—or, at least, to the classical forms of Skolem's Paradox which we have been considering—to be complete. In the next chapter, I turn to examine a new formulation of Skolem's Paradox, Putnam's so-called "model-theoretic argument against realism."

# Chapter 4

# On Putnam and his Models

> It is not my claim that the 'Löwenheim-Skolem paradox' is an antinomy *in formal logic*; but I shall argue that it *is* an antinomy, or something close to it, in *philosophy of language.* Moreover, I shall argue that the resolution of the antinomy—the only resolution that I myself can see as making sense—has profound implications for the great metaphysical dispute about realism which has always been the central dispute in the philosophy of language.
>
> Hilary Putnam: Models and Reality

For the past twenty years or so, Hilary Putnam has advanced a family of arguments which go under the patronymic "the model-theoretic argument against realism." These arguments purport to show that basic theorems of model theory entail that realistic accounts of truth and reference are untenable and that, as a result, realistic metaphysics is incoherent. On the basis of these arguments, Putnam urges philosophers to abandon traditional realism—or "metaphysical realism," as he likes to call it—and to adopt Putnam's own, new-and-improved, "internal realism" instead.

In the present chapter, I discuss three issues concerning Putnam's model-theoretic argument. First, I examine one version of the argument—a version closely connected to the traditional Löwenheim-Skolem paradox—and try to explain exactly how this version is supposed to work. Second, I show that a key step in this argument rests on an outright mathematical mistake, and I discuss some of the philosophical ramifications of this mistake. Third, I argue that, even if Putnam could get his mathematics to work, his argument would still fail on purely philosophical grounds. At the end of the day, therefore, I conclude that realists have little to fear from Putnam's model-theoretic argument.

## 4.1 Two Preliminary Clarifications

Before I begin, two clarifications are in order. The first is terminological. As Putnam uses the term "realism," it refers to a doctrine which has at least as much to do with *semantics* as it does with *ontology.* To be a "realist," one must *both* accept the existence of certain objects *and* be committed to the claim that words and phrases of ordinary English refer to these objects in a determinate manner. So, for instance, a realist about cats will believe that the world is populated by cats; she will also think that the word "cat" refers to (and only to) cats and that names like "Fluffy" and "Puffy" pick out specific cats in a determinate fashion. Similarly, a realist about set theory believes in the existence of a set-theoretic universe; she also thinks that words like "set" and "membership" refer to this universe in a determinate way.

This terminological point is important, because Putnam's model-theoretic arguments are aimed almost exclusively at the *semantic* side of "realism." Putnam doesn't argue that certain objects don't exist; instead he argues that nothing in our use of language fixes determinate reference relations between these objects and the words and phrases of ordinary English. As a result, Putnam's "anti-realist" conclusions really amount to claims of semantic indeterminacy: the word "cat" could refer to cherries, the word "mat" could refer to trees, and "set" and "membership" could refer to almost anything at all.[1]

Second, and as I noted in the introduction, Putnam has advanced several different versions of the model-theoretic argument. These versions all share certain family resemblances. They all argue against realistic accounts of truth and reference, and they all employ model theory in making their case. Nevertheless, they differ both with respect to the exact targets of their attack and with respect to the specific model-theoretic results they employ. On the target side, some versions of the model-theoretic argument focus exclusively on the semantics of mathematical English, while others widen this focus to discuss the semantics of more commonplace forms of English—e.g., ordinary talk about cats and mats, cherries and trees, flies and spiders. On the results side, some versions of the model-theoretic argument employ the Löwenheim-Skolem theorems, while others rely on theorems concerning the permutation of a model's domain.

This chapter focuses on a version of Putnam's argument which has two distinctive features: 1.) it limits itself to the case of mathematical (and, in particular, set-theoretic) language, and 2.) it uses the Löwenheim-Skolem theorems as its chief model-theoretic tool. In my view, this version of Putnam's argument is fairly

[1] It is significant, here, that these examples all involve a fairly *radical* form of semantic indeterminacy. In particular, Putnam's candidate references for words like "cat" and "set" have essentially *no* connection with independently plausible stories about the metaphysics of cats and sets. This radical indeterminacy is characteristic of Putnam's arguments and is part of what makes them so interesting.

To better appreciate this point, it's worth contrasting Putnam's arguments with arguments which make semantic indeterminacy "ride piggy-back" on genuine metaphysical questions. So, for instance, we might have a genuine question as to whether cats should be thought of as four-dimensional space-time worms or as objects which exist (in their entirety) at many different times; similarly, we might wonder whether cats should be identified with the mereological sum of their component parts or be regarded as something more substantial than these sums. Further, our difficulty in answering such questions might lead us to think that there is an underlying indeterminacy in the way we use our language: perhaps there isn't anything about our use of the word "cat" which determines whether this word refers to space-time worms or to temporally located objects, to mereological sums or to substantial individuals. In these cases, and in others like them, genuine indecision about the *metaphysics* of cats motivates a (limited) argument for the *semantic* indeterminacy of the word "cat."

Putnam's model-theoretic arguments, in contrast, do not rely on such "metaphysical motivations," and their conclusions are not limited by the need to find multiple, *equally metaphysically plausible,* candidates for the "references" of particular words. Instead, Putnam's arguments purport to show that *almost any object* can serve as the referent of *almost any word.* At one point, for instance, Putnam argues that the word "cat" could well refer to cherries and the word "cherry" to cats (see [42] ch. 2). This complete disconnect between the intuitively plausible reference(s) of our words and the reference(s) "uncovered" by Putnam's model-theoretic arguments is part of what makes these arguments so interesting (and so *infuriating*!).

representative: most of the philosophical issues raised by other versions of the argument are present here as well, and most of my criticisms of this version carry over, *mutatis mutandis*, to these other versions. In the present context, however, a full discussion of these comparative issues would take us too far afield; hence, I eschew such discussion here.[2]

## 4.2 Putnam's Argument

The version of Putnam's argument which I want to consider occurs in the first fifteen pages of a paper entitled "Models and Reality" ([43]). The paper begins with a short discussion of Skolem's Paradox and then proceeds to "spin" this paradox through three different transformations. For expository convenience, I follow this same progression in laying out Putnam's argument.

To begin, recall the generic formulation of Skolem's Paradox which appeared at the beginning of Chapter 1. We began by choosing a standard, first-order axiomatization of set theory, ZFC. We then noticed that, on the assumption that this axiomatization has a model at all, the Löwenheim-Skolem theorems entail that it has a countable model (call this model $M$). Further, we saw that Cantor's theorem—or, more properly, the fact that ZFC *proves* Cantor's theorem—entails that $M$ contains an element $\hat{m}$ such that:

$$M \models \text{“}\hat{m} \text{ is uncountable.”}$$

Finally, we noticed that, since $M$ itself is only countable, there are only countably many $m \in M$ such that $M \models m \in \hat{m}$. Hence, cardinality seems to be relative: from one perspective $\hat{m}$ seems to be uncountable, while from another perspective, $\hat{m}$ is clearly countable.

[2] That being said, let me assuage my intellectual conscience by making three brief points concerning the significance of this version of Putnam's argument. First, it is clearly the most widely *accepted* version of this argument. Critics like David Lewis and Michael Devitt have sharply criticized other versions of the argument, but they have focused on aspects of those versions which are not present in the version examined here (in particular, they have focused on those versions' treatment of non-mathematical language). More tellingly, several prominent philosophers—notably, Hartry Field and Crispin Wright—have explicitly endorsed the present version of the model-theoretic argument, while rejecting Putnam's other versions of this argument (see [17] and [65]).

Second, the fact that this version of Putnam's argument deals only with mathematical language renders it immune to challenges based on the "causal theory of reference" and/or simple demonstratives. In the case of talk about cats, we might think that causal interactions between people and cats help to constrain our theory of reference; similarly, we might think that our ability to *point* to cats while saying "that's a cat" partially determines the reference of "cats." In the case of mathematical talk, however, such considerations play no role: there aren't any causal interactions between people and $\aleph_1$, and no one *could* point to the powerset of $\aleph_{37}$. Indeed, as Paul Benacerraf has noted, the causal theory of reference poses its *own* problems for realism about mathematics (see [2]).

Third, because this version of Putnam's argument employs the Löwenheim-Skolem theorems—rather than theorems about the permutation of a model's domain—it is immune to an obvious response from mathematical structuralists. Without going into any of the details, suffice it to say that structuralists should have no objection to Putnam's permutation arguments: indeed, they may even think that these arguments *support* a certain kind of mathematical realism. On the other hand, *no* structuralist should be happy with the conclusions of Putnam's Löwenheim-Skolem argument. As a result, this version of Putnam's argument packs a stronger "punch" than other versions do.

Putnam begins his analysis of this paradox by noting that, whatever it might show about the countability or uncountability of $\hat{m}$, the paradox highlights the fact that ZFC has many *different* models.[3] Further, these models interpret many of the central definitions of classical set theory in structurally different ways. Some models satisfy "$m$ is uncountable" only if $m$ really is uncountable, while other models (like the $M$ above) satisfy "$m$ is uncountable" even when $m$ is really countable. Similarly, some models satisfy sentences like "$m$ is finite" or "$m$ is the power set of $n$" if and only if $m$ really is finite or really is the power set of $n$, while other models satisfy these sentences under (quite) different circumstances. Therefore, to the extent that the meanings of phrases like "is countable," "is finite" and "is the power set of" are fixed only by the model theory of first-order set theory, these phrases will be semantically indeterminate.

To put this point using Putnam's own terminology, Skolem's Paradox shows that it is impossible to pin down the "intended interpretation" of our set-theoretic vocabulary simply by means of (first-order) axioms. Further, Putnam thinks it is highly unlikely that anything *other* than axioms could make this situation better: "but if *axioms* cannot capture the 'intuitive notion of set' what possibly could?" ([43], 3) We have, therefore, a preliminary version of Putnam's argument:

1. Axioms cannot fix the intended interpretation of set theory.
2. Nothing other than axioms can fix the intended interpretation of set theory.

So, 3. Set-theoretic language is semantically indeterminate.

Here premise 1. follows—or is supposed to follow—from the Löwenheim-Skolem theorems, while premise 2. is just an undefended assumption. Finally, the conclusion (line 3.) is the specialization of Putnam's general anti-realism to the particular case of set theory.

Next, Putnam tries to parlay the semantic indeterminacy of set-theoretic language into more specific indeterminacies regarding the *truth-values* of set-theoretic sentences. He writes: "If I am right, then the 'relativity of set-theoretic notions' extends to a *relativity of the truth value of '$V = L$'* (and, by similar arguments, of the axiom of choice and the continuum hypothesis as well)" ([43], 7-8). Putnam's idea here is simple. It is a result of modern set theory that $V = L$ is "independent" of ZFC—that, on the assumption that ZFC itself is consistent, so are ZFC + V=L and ZFC + $V \neq L$. Hence, if the intended model of set theory is fixed *only* by the axioms of ZFC—and if there is, in fact, some such intended model—then there is an intended model in which $V = L$ is true and another one in which $V = L$ is false.[4]

Finally, Putnam tries to show that, where axiomatics fail us, physical science cannot pick up the slack. In particular, physical science cannot fix a unique "intended interpretation" for the vocabulary of set theory, nor can it restrict such interpretations so as to eliminate the indeterminacy in truth-values of set-theoretic sentences. Now, at first glance, this last argument may seem somewhat unmotivated: how, after all, *could*

[3] Assuming, of course, that ZFC has at least one model to begin with.

[4] More formally, the fact that ZFC has a model entails that ZFC is consistent. From here, the independence of $V = L$ entails that both ZFC + V=L and ZFC + $V \neq L$ are also consistent. So, by completeness, both ZFC + V=L and ZFC + $V \neq L$ have models. For more on the independence of $V = L$ (and the axiom of choice and the continuum hypothesis), see [26] or [30].

physical science affect the interpretation of basic set-theory? Later glances, however, reveal that Putnam has a reason for concern here, and that this reason is closely connected with his goal of proving that $V = L$ has an indeterminate truth-value.

Suppose that we have constructed a machine which takes a measurement—of something, it doesn't matter what—every three or four seconds. Suppose also that this machine gives a reading of 1 or 0 depending on the results of its measurements. Finally, suppose that this machine manages to run for an *infinite* period of time and (so) produce an infinite sequence of measurements. In theory, then, the sequence of ones and zeros which results from these measurements could "code up" some non-constructable set—i.e., some set which lives in $V$ but not in $L$. In this case, it might seem like *nature itself* has managed to falsify the hypothesis that $V = L$. This possibility is why Putnam thinks he needs an explicit argument regarding the impact of physical science on the interpretation of set theory (at least, that is, if he wants to continue to maintain his claim that sentences like "$V = L$" have indeterminate truth values).

Putnam's explicit argument involves modifying and defending each of the two premises in the "preliminary" argument sketched a moment ago. He begins with the first premise of this argument, and modifies it to care of the quasi-empirical concerns mentioned in the last paragraph. His principle tool here is the following extension of the Löwenheim-Skolem theorem (see, [43] 6):

**Theorem 1:** *ZF plus V=L has an $\omega$-model which contains any given countable set of real numbers.*

Given this theorem, Putnam argues as follows. Let OP be a countable collection of real numbers which codes up all of the measurements human beings will ever make. By Theorem 1, there is a model of *ZF + V=L* which contains OP (or, at least, a formal analog of OP). Since this model satisfies ZFC it must be an "intended model"; since it both satisfies $V = L$ and contains OP, it takes care of the measurement problem from the last paragraph. Hence, Putnam's problem with physical science seems to be solved.

Further, Putnam thinks this particular problem is the *only* problem with physical science that *needs* to be solved—i.e., the only problem that might impede his project of finding a model which both takes physical science into account and satisfies the axiom $V = L$. Putnam writes:

> Now, suppose we formalize *the entire language of science* within the set theory ZF *plus* $V = L$. Any model for ZF which contains an abstract set isomorphic to OP can be extended to a model for this formalized language of science which is *standard with respect to OP*; hence... we can find a model *for the entire language of science* which satisfies '*everything is constructable*' and which assigns the correct value to all physical magnitudes ([43] 7).

Thus, as long as the only constraints on the interpretation of our set-theoretic vocabulary come from the formal structure of our scientific theories (including the explicit axioms of our set theory) and from the physical measurements we happen to make, there will be *an* interpretation of set theory on which $V = L$ comes out true. On the flip side, Putnam simply assumes that there will be *some* interpretation of our set theory—again, an interpretation compatible with the rest of our scientific theories and with all the physical measurements we might ever happen to make—on which $V = L$ comes out false.[5]

[5] Putnam makes this assumption for two reasons. From a dialectical perspective, Putnam takes himself to be arguing

This, then, gives us a revision of the first premise in Putnam's "preliminary" argument. If we let "theoretical constraints" refer to the set of sentences which provides our best theory of the world—i.e., our best physical theory together with our standard axioms for set theory—and if we let "operational constraints" refer to all the measurements we might ever happen to make, then we obtain:

$1'$. Theoretical and operational constraints do not fix a unique "intended interpretation" for the language of set theory.

Putnam also accepts the obvious analog of premise 2. in our "preliminary" argument. His reasons for accepting this premise will be discussed in section 4.4. For now, let me simply lay out the remainder of his argument:

$2'$. Nothing other than theoretical and operational constraints *could* fix a unique "intended interpretation" for the language of set theory.

So, $3'$. There is no unique "intended interpretation" for the language of set theory.

Finally, because different, equally "intended," interpretations of set theory disagree on the truth value of $V = L$, there simply is no determinate truth value for this sentence: it is, in Putnam's words, "just true in some intended models and false in others" ([43] 5).

This, then, is the overall structure of Putnam's argument. The goal is to show that set-theoretic language is semantically indeterminate. To reach this goal, we first note that the axioms of set theory do not determine a unique interpretation for set-theoretic language. Next, we observe that throwing in scientific information—e.g., the physical theories and measurements which fill out Putnam's "theoretical and operational constraints"—does not improve the situation. Hence, to the extent that the "intended interpretation" of set-theoretic language is *determined* by theoretical and operational constraints, there will be many different "intended interpretations" of set theory. Finally, we notice that these different "intended interpretations" disagree on the truth value of sentences like $V = L$, and we conclude that these truth values are, themselves, indeterminate.

## 4.3 The Mathematics of Premise 1

In this section, I slow down and examine more carefully the details of Putnam's mathematics. My main theses are straightforward. I argue 1.) that Putnam's proof of Theorem 1 is mistaken, 2.) that this mistake cannot be "patched up" without weakening the overall force of Putnam's argument, and 3.) that even the weakened version of Putnam's argument leaves the realist with some significant problems.

against Gödel. Gödel thought there *was* a unique "intended interpretation" of set theory and that $V = L$ was false on this interpretation. Hence, the existence of an intended model satisfying $V \neq L$ isn't really at issue in this context.

From a somewhat deeper perspective, it's fairly easy to start with a model which *does* satisfy $V = L$ and expand it to one which *does not* (and to preserve nice properties like "being an $\omega$-model" in the process). It's substantially more difficult to start with a model which *does not* satisfy $V = L$ and expand it to one which *does*. This is undoubtedly the main reason Putnam puts so much effort into obtaining a model which satisfies $V = L$ and so little into obtaining one which satisfies $V \neq L$.

To begin, consider Putnam's proof of Theorem 1. The theorem says that if $X$ is a countable collection of real numbers, then there exists an $\omega$-model, $M$, such that $M \models ZF{+}V{=}L$, and $M$ contains an "abstract copy" of the set $X$. Putnam's proof begins by noting that, in the special case in which we allow $M$ to be countable, we can code both $M$ and $X$ by single reals. In this case, the theorem can be formulated as a $\Pi_2$ sentence of the form: (For every real $s$) (There is a real M) such that $(\ldots M, s, \ldots)$. From here, Putnam argues as follows:

> Consider this sentence *in the inner model* $V = L$. For every *$s$ in the inner model*—i.e., for every $s$ in $L$—there is a model—namely $L$ itself—which satisfies '$V = L$' and contains $s$. By the downward Löwenheim-Skolem theorem, there is a countable submodel which is elementarily equivalent to $L$ and contains $s$. (Strictly speaking, we need here not just the downward Löwenheim-Skolem theorem, but the 'Skolem Hull' construction which is used to prove that theorem.) By Gödel's work, this countable submodel itself lies in $L$, and, as is easily verified, so does the real that codes it. So, the above $\Pi_2$-sentence is true in the inner model $V = L$.
>
> But Shoenfield has proved that $\Pi_2$-sentences are *absolute*: if a $\Pi_2$-sentence is true in $L$, then it must be true in $V$. So the above sentence is true in $V$. ([43] 6)

Ironically enough, the problem with this "proof" involves Putnam's application of the Löwenheim-Skolem theorem. The "short version" of the problem is quite simple: the downward Löwenheim-Skolem theorem applies only to structures which have sets for their domains, and $L$—the structure to which Putnam applies the downward Löwenheim-Skolem theorem—doesn't have a set for its domain. Hence, Putnam cannot (legitimately) use the downward Löwenheim-Skolem theorem to obtain "a countable submodel which is elementarily equivalent to $L$ and contains [the real] $s$." Without this submodel, Putnam cannot ensure that $L$ satisfies the $\Pi_2$ sentence which he needs it to satisfy; hence, he cannot apply Shoenfield absoluteness to "reflect" this sentence up to $V$. In the absence of this countable submodel, therefore, Putnam's overall proof simply collapses.[6]

[6] The "long version" of this problem simply expands the "short version" with a fair bit of terminological clarification. To begin, set theorists typically distinguish between *sets* and *proper classes*. Roughly, sets are classes which are "small enough" to be members of other classes, while proper classes are classes that are "too big" to count as sets. Examples of sets would include $\emptyset$, $\aleph_0$, and the power set of $\mathbb{N}$. Examples of proper classes would include the class of all sets, the class of all ordinals, and the class of all countable sets. Most pertinently, the class of all constructable sets—i.e., $L$—is a proper class.

It is important here, for reasons which will become clear shortly, that that proper classes are usually required be *definable*. That is, for some formula in the language of set theory, $\phi(x, y_1, \ldots, y_n)$, and some sequence, $a_1, \ldots, a_n$, of sets, we can consider the class of all sets $b$ such that $\phi(b, a_1, \ldots, a_n)$ is true. To insist that proper classes be definable is to insist that *all* classes be picked out by a formula of this kind.

Second, we need to distinguish between two interpretations of the terms "model" and "satisfaction." In ordinary model theory, the term "model" refers exclusively to structures which have *sets* for their domains. In turn, the "satisfaction" relation is defined as a relation *between sets*—i.e., between the sets which constitute the domain (and relations) of a model and those which code up the formulas of our language.

Set theorists often use the terms "model" and "satisfaction" somewhat differently. They often speak of "class models" when they want to refer to proper classes in which certain collections of sentences hold, and they often use "satisfaction" to refer to the fact that certain sentences become true when their quantifiers are relativized to a proper class—i.e., when they are explicitly relativized using the formula which defines the class in question (see above). It is in this latter sense, for instance, that Putnam refers to "inner models."

### 4.3.1 Five Technical Comments

Before I discuss the philosophical ramifications of the failure of Putnam's proof, I want to make five technical comments about the failure itself. First, Putnam's proof is not saved by his qualification: "Strictly speaking, we need here not just the downward Löwenheim-Skolem theorem, but the 'Skolem Hull' construction which is used to prove that theorem." To be sure, the Skolem Hull construction allows us to prove so-called "reflection" theorems in which some *finite* collection of sentences is "reflected" from a proper class to a set. That is, if we start with a proper class which "satisfies" some finite collection of sentences, then the Skolem Hull construction lets us find a countable set which satisfies the same collection of sentences.[7] However, this construction *only* works when when we try to reflect *finite* collections of sentences. In particular, then, it does not allow Putnam to reflect the *infinite* collection, $ZF + V = L$.

Second, there is nothing particularly surprising about the fact that Putnam's proof fails. Leaving aside the details of this proof, consider just the *form* of Theorem 1: for any countable set of reals $X$, there is an $\omega$-model $M$, such that $M \models$ *ZF+V=L* and $X \in M$. Now, since any model of *ZF + V=L* is also a model of ZFC, Theorem 1 entails that there is a model for ZFC. And since this, in turn, entails that ZFC is consistent, Theorem 1 also entails that ZFC is consistent.

However, by Gödel's second incompleteness theorem, the consistency of ZFC *cannot* be proved from within ZFC itself (assuming ZFC is consistent). As a corollary, then, Theorem 1 cannot be proved in ZFC. Therefore, we should not be surprised to find that there is a mistake in Putnam's proof: Putnam's proof must be mistaken, because Theorem 1 *can't* be proved using the set theory with which Putnam is working. Unless Putnam is willing to adopt some stronger set theory, his overall argument is bound to fail.[8]

For our purposes, the key fact concerning all this terminology is the following: the Löwenheim-Skolem theorems apply to set models, but they do not apply to class models. Hence, we cannot use the downward Löwenheim-Skolem theorem to find an elementary submodel of some class model of ZF. In particular, Putnam cannot use the downward Löwenheim-Skolem theorem to find a (countable) model which is elementarily equivalent to $L$.

Given this, Putnam faces a dilemma. On the one hand, if he intends the term "model" to refer to set models, then his proof goes wrong when he says: "for every $s$ in $L$—there is a model—namely $L$ itself—which satisfies '$V = L$'." For, since $L$ is not a model at all, it is not a model satisfying $ZF + V = L$. On the other hand, if Putnam intends the term "model" to refer both to set models and to class models, then he can legitimately call $L$ a "model." But even then, *he cannot apply the downward Löwenheim-Skolem theorem to this model.* On either hand, then, Putnam's proof is mistaken. Whichever way the term "model" gets defined, Putnam's proof requires that we apply the downward Löwenheim-Skolem theorem to $L$; as we have seen, this is just not possible.

[7] Here, we have an example of a *legitimate* shift between the two uses of "satisfies" that I mentioned in the last footnote.

[8] In conversation, several people have suggested that Putnam might not be working in ZFC in the first place—i.e., that he might *already* be working in some stronger form of set theory. But, while this suggestion helps to save Putnam's mathematics, it faces two major difficulties. The first is textual: Putnam's entire paper is *about* ZFC, and throughout the paper Putnam writes as though ZFC is the *obvious* set theory in which to work. Hence, it would be strange to find that Putnam's entire argument rests on an (utterly unmentioned) version of set theory which Putnam keeps hidden in the background. The second difficulty is philosophical: for reasons which will be discussed shortly, working with a stronger set theory would not, in the long run, save Putnam's argument (although it *might,* in the short run, allow Theorem 1 to be proved properly).

Third, Putnam's proof is relatively easy to "patch up." If Putnam *is* willing to adopt a stronger set theory, then he can easily salvage his theorem (and, for the most part, his proof). So, for instance, if Putnam were willing to accept the existence of inaccessible cardinals, then his proof could be reconstituted with only minor modifications.[9] Similarly, if Putnam were willing to extend ZFC with a collection of axioms governing the behavior of proper classes, then he could probably prove "full reflection theorems" which would allow him to obtain (elementary) submodels of proper classes like $V$ and $L$; this would, once again, allow him to reconstitute the essentials of his original proof. In either case, therefore, a slight strengthening of Putnam's background mathematics allows Putnam to save his proof from the problem discussed above.

Fourth, although these "patching strategies" allow Putnam to salvage his proof, they do very little towards salvaging his overall philosophical argument. Very roughly, patchings of the type just mentioned leave Putnam's critics with two fairly obvious lines of response. On the one hand, some philosophers might reject Putnam's argument *simply* because they reject his new mathematics. Whereas ZFC is a relatively widely-accepted axiomatization of set theory, the extensions discussed above are less-widely accepted. Hence, a philosopher (or mathematician) who has reservations about inaccessible cardinals and/or strong class axioms might well reject Putnam's new argument just because it employs this extra mathematics. Indeed, such a philosopher might even think that she has been given new *reasons* for rejecting Putnam's extra mathematics: "if Putnam's new math generates arguments for semantic indeterminacy, then so much the worse for Putnam's new math."

On the other hand, even philosophers who *do* accept Putnam's strengthened mathematics—say, those who accept the axiom of inaccessible cardinals—have ample grounds for rejecting Putnam's overall argument. In particular, they should reject Putnam's claim that his model $M$—the model guaranteed by Theorem 1—satisfies "all theoretical constraints." Since Theorem 1 does not guarantee that $M$ satisfies the sentence "there exists an inaccessible cardinal," $M$ may not even satisfy the "theoretical constraints" imposed by *set theory.* Contra Putnam, then, there is no reason to think that $M$ provides an "intended interpretation" for set-theoretic language.

Recall, here, the philosophical *point* of Putnam's theorem. Putnam wants a model of $ZF + V{=}L$ which "satisfies all theoretical constraints... [and] all operational constraints as well" ([43] 7). His theorem is supposed to provide such a model (where the theory of the model takes care of "theoretical constraints" and some countable collection of real numbers takes care of "operational constraints"). With this model in hand, Putnam tries to argue that $V = L$ is true in the (or at least in *an*) intended model of set theory.

My point is simply this: if our working set theory is *stronger* than ZFC—because we accept inaccessible cardinals, or class axioms, or whatever else is needed to patch up Putnam's proof—then it's hard to see how Putnam's theorem accomplishes its goal. Given that we accept more mathematics than ZFC, this new

[9] To see this, let $\kappa$ be the inaccessible cardinal in question. At the point in Putnam's proof where he claims that $L$ is a model for $ZF + V{=}L$ and that $L$ contains the real $s$, Putnam can simply argue that $L_\kappa$ is a model for $ZF + V{=}L$ and that $L_\kappa$ contains the real $s$. Since $L_\kappa$ *really is* a set model for $ZF + V{=}L$, Putnam can proceed to apply the downward Löwenheim-Skolem theorem to obtain his desired model $M$. From here, Putnam's proof proceeds just as before.

mathematics should count as part of our "theoretical constraints." Thus, it's not enough for Putnam to build a model which satisfies *ZF + V=L;* Putnam needs a model which satisfies *ZF + V=L plus whatever else we happen to have added to* ZFC. Since Putnam's theorem does not, so far as we know, provide a model satisfying this *extended* theory, it doesn't do what Putnam wants it to do.

Fifth, this problem is intrinsic to the *kind* of argument Putnam wants to make. Returning again to the incompleteness theorem, we note that there is no way for Putnam to both 1.) use a particular axiomatization of set theory (say, ZFC + XYZ) as his background set theory and 2.) prove the existence of a model satisfying ZFC + XYZ +V=L. Hence, *whatever* version of set theory Putnam winds up working in, he will be unable to tailor his overall argument so as to take care of the "theoretical constraints" this version imposes.

As a result, Putnam faces an inescapable dilemma. If he pitches his argument towards philosophers who accept *less* set theory than he himself does, then these philosophers will reject his argument *simply because they reject the set theory used in proving Putnam's key theorem.* If he pitches his argument towards philosophers who accept *the same* set theory that Putnam does, then his argument can't take care of *these philosopher's* "theoretical constraints." At the end of the day, therefore, *no philosophers* will have adequate grounds for accepting Putnam's model-theoretic argument.[10]

This, therefore, gives us a first, and essentially technical, response to Putnam's argument. Putnam's argument depends on a key theorem which Putnam is not in a position to prove. Nor, for reasons relating to Gödel's second incompleteness theorem, can he place himself in such a position without jeopardizing the very philosophical point which his theorem is supposed to support. In short: the mathematical mistake which we discussed at the beginning of this section is one which cannot be fixed without undercutting the model-theoretic argument as a whole.

### 4.3.2 Two Philosophical Comments

How strong is this argument which I have just sketched? Unfortunately, much as I like the argument, I'm afraid the answer is: "not very." On the positive side, the argument shows that Putnam cannot *conclusively prove* that set theory is semantically indeterminate. That is, if we want Putnam to stand toe-to-toe with the realist and *prove* that semantic anti-realism is true, then the argument of the last section shows that he cannot do it.[11]

In addition, the argument provides what I like to call the "cocktail party" response to the skeptic. Here, the realist simply demands that Putnam prove that "unintended interpretations" of set theory really exist. If Putnam uses some non-ZFC mathematics to give his proof, the realist professes to disbelieve this new mathematics. If Putnam then gives *reasons* for accepting the new mathematics, the realist readily accepts

[10] Note that this argument is not specific to the search for a model satisfying $V = L$; it works equally well against attempts at finding models for ZFC $\pm$ CH or ZF $\pm C$. Indeed, when framed as a simple incompleteness problem, it even applies to the "plain vanilla" version of Putnam's argument which merely tries to obtain structurally dissimilar models of ZF.

[11] At least, not using the kind of argument which we have so-far been examining.

those reasons (and the new mathematics), but she then questions whether her (newly acquired) "theoretical constraints" have really been met. At no point in the resulting dialectic will Putnam obtain the upper hand, so our realist can leave the party in relatively good spirits.

Unfortunately, this "cocktail party" response is better suited to APA smokers than to a serious analysis of the model-theoretic argument. For, even if Putnam can't get himself into the right dialectical position to *prove* semantic indeterminacy, he has clearly given the right kinds of arguments to raise the *possibility* of such indeterminacy. Suppose that premise $2'$ in Putnam's argument is correct. Then the only way for set-theoretic language to be semantically determinate is for there to be a *unique* model satisfying all of our "theoretical and operational constraints." To the extent that other such models *happen to exist,* set theory winds up being semantically indeterminate.

In this context, then, the mere fact that Putnam cannot *prove* his central theorem—or whatever extensions of this theorem are needed to take care of *our* theoretical constraints—provides very little comfort. If the technical response to Putnam's argument is all we have to go on—if, that is, we are willing to accept premise $2'$ and to rest our challenge to premise $1'$ solely on the considerations discussed above—then realism depends on the *mere hope* that Putnam's non-standard models don't happen to exist. This not much to stake an entire metaphysics on![12]

All that being said, this argument clearly depends on the assumption that premise $2'$ is true. If it not, then the existence of non-standard models for our "theoretical and operational constraints" is considerably less troubling. It is high time, therefore, that we turn and examine premise $2'$.

## 4.4 The Philosophy of Premise 2

Although premise $1'$ of Putnam's model-theoretic argument has attracted a great deal of attention in the literature, it should be fairly clear that premise $2'$ is where the real philosophical action takes place. For one thing, premise $1'$ only eliminates one possible method for fixing the intended interpretation of set theory; premise $2'$ tries to eliminate all other methods in one fell swoop. For another thing, most of the methods which have actually been proposed for fixing the intended interpretation of our language—e.g., causation in the case of other versions of the model-theoretic argument and set-theoretic "perception" in the case of the present version (see, [11] and [32])—clearly fall under premise $2'$ rather than premise $1'$. Hence, $2'$ is the premise which warrants the most sustained philosophical attention.

In this section, I consider three arguments which Putnam might use to defend premise $2'$. The first two arguments have played only a minor role in Putnam's own discussions of the model-theoretic argument, but I think they shed a great deal of light on that argument's overall structure. In particular, I think they

[12] Note, for instance, that we have to hope that there are no large cardinals which are stronger than the ones explicitly mentioned in our "theoretical constraints." Leaving large cardinals aside, we have to hope that a great many small models of set theory just happen to be left out of the set-theoretic universe. And, while this is certainly a *consistent* hope, that's about all that can be said for it.

highlight, if only through their failures, exactly what Putnam needs to do if he wants to give a philosophically substantial defense of premise $2'$. The third argument—the so-called "just more theory" argument—*has* played a large role in Putnam's own writings (and has received an *enormous* amount of attention in the recent literature). In 4.4.3–4.4.4, I explain what's wrong with this third argument, and in 4.5, I discuss some connections between the failure of this argument and the more technical failures discussed in section 4.3.

### 4.4.1 The No-Explanation Argument

The first argument—which I shall call the "no-explanation" argument—is Putnam's most straightforward defense of premise $2'$. Although the argument does not appear in Putnam's *original* development of the model-theoretic argument, Putnam has recently started to use it as something something of a fall-back position. Further, this argument is regarded as the *key* to the model-theoretic argument by at least one (friendly) commentator on Putnam.[13]

The no-explanation argument rests on two observations. First, realists need a plausible account of how mathematical language refers to mathematical objects. If we maintain that the truth of "$\aleph_0 < \aleph_1$" *depends on* the fact that $\aleph_0$ is smaller than $\aleph_1$, then we need an account of how "$\aleph_0$" is related to $\aleph_0$ and "$\aleph_1$" is related to $\aleph_1$. Indeed, even if we simply maintain that there is an objective fact of the matter concerning the truth-value of "$\aleph_0 < \aleph_1$," we need some story to explain this objectivity.

Second, realists have yet to provide a plausible account of how reference to mathematical objects is supposed to work. Although such an account is necessary for realism, and although this necessity has been obvious for some time, no attractive candidates (at least by Putnam's lights) have yet been put forward. Hence, until such candidates are forthcoming, we should give (at least provisional) support to premise $2'$. In short: the need for an explanation of mathematical reference, combined with the lack of a good suggestion as to how such an explanation might go, leads to the thought that no explanation is possible.[14]

[13] See [45] for Putnam's use of this argument as a fall-back position. See [1] for a recent discussion of the argument and for a defense of the view that this argument is central to Putnam's overall project.

[14] It is worth noting that the no-explanation argument can be *supplemented* by more detailed arguments against specific realistic proposals. In Putnam's case, this supplementation comes in two varieties: insults and substantial arguments against the causal theory of reference. On the insults side, Putnam gets a fair bit of mileage out of sounding incredulous when various reference-fixing mechanisms are proposed. Platonistic accounts of "grasping concepts" and "intuiting mathematical objects" are rejected as "unhelpful as epistemology and unpersuasive as science" ([43] 10). Similarly, Chisholm's account of intentionality is rejected as "unhelpful epistemology and almost certainly bad science as well" ([43] 14). Finally, David Lewis' appeal to the theory of universals is dismissed as "medieval" ([44] xii), and causal theories of reference are derided as follows:

> In the context of a twentieth-century world view, to say in one's most intimidating tone of voice 'I believe that causal connections determine what our words correspond to' is only to say that one believes in a *one-knows-not-what* which solves our problem *one-knows-not-how.* ([44] xii )

On the more substantial side, Putnam has presented a number of detailed arguments against the causal theory of reference (see [40], [44] essays 11–13, and [45]). If these arguments were correct, then they would show that *one particular* mechanism for fixing the reference of our language does not work. However, since the causal theory is largely irrelevant to the (mathematical) languages at issue in this chapter, and since ruling out a *particular* account of reference doesn't really suffice to establish $2'$, I will say no more about these (admittedly more substantial) arguments here.

At it heart, this no-explanation argument amounts to a (relatively well-motivated) case of burden-juggling. Putnam claims that it's up to realists to provide an adequate semantics for mathematical language. Until realists can meet this burden, we should default to the view that mathematical language has no fixed semantics (or, at least, that its semantics are *no more* fixed then "theoretical and operational constraints" can ensure). What's more, realists' historical failure to provide an account of mathematical language which even comes close to meeting this burden, should lead us to think that our "default" position is actually pretty solid.

Unfortunately, this kind of burden-juggling suffers from two fairly obvious flaws. First, it really amounts to more of a theoretical challenge to realists than to a genuine *argument* against realism. Although the no-explanation argument highlights a particular problem which realists need to solve—and notes that this problem has not *yet* been solved—the argument does nothing towards showing that the problem *can't* be solved. That is, the argument gives no positive reasons for thinking that realists can't, in the long run, explain the mechanisms that fix the interpretation of set theory, and it certainly gives no reasons for thinking that such mechanisms fail to exist.

It is important, here, to recall the strength of the claims that Putnam's model-theoretic argument is supposed to defend. Putnam's argument is intended as a defense of claims like the following:

> The idea that it is something *other* than operational and theoretical constraints that singles out the right reference relation . . . is an *incoherent* idea. ([45], 215)
>
> The supposition that even an 'ideal' theory (from a pragmatic point of view) might *really* be false appears to collapse into *unintelligibility.* ([41], 126)
>
> The 'Löwenheim-Skolem paradox' *is* an antinomy, or something close to it, in *philosophy of language.* ([43], 1)

The no-explanation argument, however, doesn't support all this talk of "incoherence," "unintelligibility" and "antinomy." At best, it supports talk of "puzzles yet to be solved" and "phenomena yet to be explained." This talk, however, is very different (and substantially weaker) than the talk which Putnam's model-theoretic argument was originally supposed to defend.

Second, the no-explanation argument is not even remotely model-theoretic. No model-theoretic insight is involved in either of the two observations on which the argument depends. Nor does model theory play any role in explaining how these observations provide *prima facie* evidence for $2'$. Thus, the no-explanation argument can be presented in its entirety without invoking any model theory. As Putnam's overall argument is supposed to be deeply model-theoretic, this lack of model-theoretic entanglement is somewhat embarrassing.

Of course, Putnam's argument for $1'$ is still model-theoretic. But, as I have already noted, the argument for $2'$ is where the real meat of Putnam's argument has to lie; hence, using model theory in defense of $1'$ is simply not enough. This is particularly true given that no one has ever adopted the kind of "implicit definition" account of reference-fixing which $1'$ purportedly rules out. If no one accepts the account of reference-fixing which Putnam's model theory rules out, and if model theory plays no role in eliminating any other accounts of reference-fixing, then it's hard to see why Putnam's overall argument should really count as "model-theoretic."

The situation here can be compared to one in which a philosopher gives an "astronomical argument against realism." After some careful work with his telescope, this philosopher claims to have refuted the hypothesis that little green men from Mars fix the reference of human language.[15] Good astronomy, after all, shows that martians don't exist! This philosopher could then go on to endorse the obvious analog of premise $2'$ in Putnam's argument: "nothing other than martians could possibly fix the reference of human language." If challenged on this premise, our philosopher could defend it by appealing to Putnam's own no-explanation argument.[16]

Clearly, however, the resulting argument would not be attractive to philosophers. Even more clearly, it wouldn't deserve the name "the *astronomical* argument against realism." The astronomy in this argument serves only to eliminate a hypothesis which no one has ever taken seriously (or even proposed as a joke!); all the philosophical work takes place in a completely non-astronomical portion of the argument. The same, then, holds of Putnam's argument. Since no one has ever proposed the account of reference-fixing that premise $1'$ rules out, the philosophical work in Putnam's argument must live entirely in his defense of $2'$. If Putnam defends $2'$ using the no-explanation argument, then his overall argument will wind up being no more model-theoretic than the argument above was astronomical.[17]

To summarize: the no-explanation argument is simply a philosophical optical illusion. It purports to provide a model-theoretic argument against the very coherence of realism. It actually provides *no* argument against realism *per se,* and it uses model theory only to refute a proposal which no one ever has (or, in my view, was ever likely to) put forward. Hence, if the no-explanation argument is the only defense of premise $2'$ that Putnam has to offer, then Putnam's overall model-theoretic argument fails—and fails rather spectacularly—to live up to its advance billing.

All that being said, the no-explanation argument does have one real virtue: it makes explicit just what Putnam needs to do to if he wants to give a genuine defense of premise $2'$. First, he need an argument which provides positive reasons for thinking that realists *can't* explain the mechanisms which fix the reference of mathematical language (or, even better, an argument which shows outright that *no such mechanisms exist*). Second, he needs to ensure that this argument makes intrinsic use of model theory (to avoid being classed with the "astronomical argument" above). Only if Putnam can meet these two conditions will he have a defense of premise $2'$ which allows his overall model-theoretic argument to work "as advertised."

[15] For the purposes of this example, it doesn't matter *how* the martians are supposed to do this. Perhaps they look very quickly from the people using a bit of language to the things to which that language refers. Perhaps they have direct contact with the denizens of Frege's third realm, and they use this contact to help us in our linguistic endeavors.

[16] The no-explanation argument, after all, says nothing about the particular hypothesis ruled out by premise $1'$. It simply notes that none of the stories about reference-fixing which realists have actually proposed have managed to stand up to careful philosophical scrutiny. Hence, it can be deployed in conjunction with almost any variation on Putnam's $1'$.

[17] Actually, given the similarities in these two arguments' overall structure, I think they should probably be taken equally seriously. I don't, however, find the astronomical argument compelling.

### 4.4.2 If axioms can't...

> But if axioms can't fix the intended interpretation of set theory, what possibly could?
>
> Hilary Putnam: Models and Reality

A second way that Putnam might argue for premise $2'$ is through reflection on the role that axioms play in formalizing and clarifying mathematical notions. Very roughly, Putnam could try to show that there is something about the way mathematicians use axioms which makes the kind of "implicit definition" at issue in premise $1'$ the most likely candidate for fixing the intended interpretation of set-theoretic language. He could then use the relative unlikeliness of other candidates to give some support to premise $2'$.

Since I have already discussed the role of axioms in mathematics in some detail in section 3.2, and since Putnam does not elaborate on his own reasons for focusing on axioms as much as he does, I will avoid a full rehash of these matters here.[18] Instead, I will simply highlight two reasons why this *kind* of argument might prove attractive to Putnam. I will then examine what I take to be the most pressing problem for this kind of argument in the context of Putnam's overall project.

To begin, one reason that Putnam might like this kind of argument is because it clearly meets the two conditions laid down at the end of the last section. If Putnam can show that treating axioms as implicit definitions is the *best* way to fix the intended interpretation of set-theoretic language, then he will have satisfied condition 1. Further, since any discussion of implicit definitions will have to focus fairly heavily on model theory—e.g., in giving its analysis of *what* implicit definitions are and *how* they work—condition 2 will be satisfied as well.

A second reason that Putnam might like this kind of argument is because it has the potential to persuade philosophers who remain unpersuaded by other versions of Putnam's argument. In particular, philosophers who object to Putnam's extension of the model-theoretic argument to the case of non-mathematical language might find the present defense of premise $2'$ somewhat reassuring. If the best reasons for accepting $2'$ involve reflection on the role of axioms *in mathematics*, then it's fairly clear why these reasons don't generalize to non-mathematical cases. Even if there are strong similarities between the ways reference gets fixed in the mathematical case and the ways it gets fixed in non-mathematical cases, these similarities will not involve the (unusually large) role which axioms play in mathematics.[19]

These, then, are two reasons why a focus on axioms might prove attractive to Putnam. Unfortunately, however, there are serious difficulties with the idea that the axioms of set theory serve to provide "implicit definitions" for the vocabulary of set theory (see 3.2.2–3.2.4). Further, at least one of these difficulties highlights a real problem for Putnam's overall project (i.e., not only for the "if axioms can't..." defense of

[18] Readers who may desire such a discussion should consult section 3.2 directly.

[19] I'm not sure that this second point sheds much light on Putnam's *own* reasons for discussing the role of axioms in mathematics. Clearly, the project of distinguishing different versions of the model-theoretic argument (with the express goal of *rejecting* at least some of them!) will not prove attractive to Putnam. Nevertheless, it will prove attractive to many other philosophers, so it warrants at least some mention here.

premise $2'$, but also for the more general line of reasoning in the model-theoretic argument as a whole). To get this problem on the table, we need to begin by recalling just how many different kinds of background semantics can be used to provide a “model-theory” for the axioms of set theory.

For convenience—and since I have already discussed some of this material in 3.2.4—I will focus here only on three broad *classes* of background semantics: trivial semantics, rich semantics and standard first-order semantics. A trivial semantics connects axioms to models in a manner which preserves little, if any, of our ordinary understanding of what the axioms in question mean (and, more importantly, of *how* they manage to mean this). So, for instance, suppose that $\mathbb{M}$ is a model and that $\Gamma$ is a collection of sentences. Then we can define a relation $\models_{\mathbb{M},\Gamma}$ such that for any model $\mathbb{N}$ and any sentence $\phi$,

$$\mathbb{N} \models_{\mathbb{M},\Gamma} \phi \Longleftrightarrow \mathbb{N} = \mathbb{M} \text{ and } \phi \in \Gamma.$$

Given this definition, $\mathbb{M}$ is the only model which “satisfies” any sentences, and $\mathbb{M}$ “satisfies” exactly the sentences which happen to be in $\Gamma$. Since $\mathbb{M}$ and $\Gamma$ are completely arbitrary, this amounts to a trivialization of the notion of “satisfaction.” After all, the reason why $\mathbb{M} \models_{\mathbb{M},\Gamma} \Gamma$ has nothing to do either with the structure of $\mathbb{M}$ or with the particular sentences in $\Gamma$; instead, the definition of $\models_{\mathbb{M},\Gamma}$ just *stipulates* that $\mathbb{M}$ “satisfies” $\Gamma$. A similar point could be made about the supermodel construction from section 3.2.3.

In contrast, a rich semantics connects axioms to models in a manner which preserves much, if not all, of our ordinary understanding of what the axioms in question mean (and how they mean it). The simplest example here would be the semantics of a “fully-interpreted language” in which phrases like “is a set” and “is a member of” are connected to the world *through* their ordinary English meanings; hence, sentences come out *true at the world* just in case they are really *true.* For a somewhat less rich example, we might consider the semantics of second-order set theory, in which the satisfaction relation requires that second-order quantifiers range over the *full* power set of the domain of any given model.

Finally, we can consider the case of first-order semantics. Here, our conception of satisfaction requires that the symbols $\neg$, $\rightarrow$, $=$, and $\exists$ have their ordinary significance, but it allows other symbols to vary quite wildly as we move from model to model. Hence, first-order semantics is intermediate between the trivial semantics discussed two paragraphs ago and the rich semantics discussed in the last paragraph: it doesn’t make the satisfaction relation as arbitrary as $\models_{\mathbb{M},\Gamma}$ (or as weak as $\models_s$), but neither does it insist that this relation “go right” about as many things as the second-order satisfaction relation (or the relation associated with a “fully-interpreted language”).

These, then, are three broad classes of background semantics under which we can interpret the axioms of set theory. To keep the relevant examples handy, I provide the following table:

| Trivial | First Order | Strong |
|---|---|---|
| $\models_{\mathbb{M},\Gamma}$ | | $\models_{\text{fin}}$ |
| $\models_s$ | $\models_1$ | $\models_2$ |
| $\models_{\text{true}}$ | | $\models_{\text{sets}}$ |

With these three kinds of background semantics in mind, I want to make four points about the role axioms play in mathematics.[20] First, axioms pick out models only against the backdrop of a fixed semantics under which those axioms are interpreted. This involves at least two things: 1.) a specific understanding of what counts as a "model" for the language of our axioms and 2.) a specific "satisfaction" relation which determines when a particular "model" can be said to "satisfy" a particular axiom. Only when these things are in place does it make sense to talk about the collection of "models" which a particular set of axioms "picks out."

Second, the axioms used in mathematics often assume a rich background semantics. So, for instance, the axioms used to define topological spaces and homology theories assume a fairly complicated second-order semantics. The axioms used in analysis assume that we can use notions like "order-complete" with their ordinary significance (e.g., to pick out the class of linear continuums). Even the classical axioms for arithmetic and set theory—i.e., the original second-order Peano axioms and Zermelo's second-order formulation of ZFC—assume a second-order semantics in the background.

Third, Putnam's argument assumes that the only way axioms can pick out their own models is through some form of *first-order* implicit definition. From a structural standpoint, this assumption is needed to make Putnam's defense of $2'$ "mesh" with his defense of $1'$. Since his defense of $1'$ rests on *first-order* model theory, his argument for $2'$ has to rule out all other forms of model theory—e.g., second-order model theory—as legitimate means of fixing the intended interpretation of set-theoretic language. Thus, if Putnam's argument for $2'$ focuses on the role axioms play in clarifying our mathematical concepts, then part of what he has to show is that this clarification only involves *first-order* semantics.

From a more textual standpoint, this assumption is fairly explicit in Putnam's own presentation of the model-theoretic argument. In some places (especially [43] and [45]) Putnam simply starts with the assumption that we have "formalized" our mathematical and scientific theories and then characterizes this "formalization" in terms of *recasting* our theories in first-order notation. In other places, Putnam considers non-first-order theories and argues that these theories don't impede his project *because they can be reinterpreted so as to take on first-order semantics.* Examples of this latter maneuver can be found at, for instance, [43] pp. 8–9 (for the case of language involving counterfactuals) and [43] p. 23 (for the case of second-order languages). Thus, not only is the assumption that axioms must (really) be first-order an assumption which is necessary for Putnam's argument, it is also an assumption which Putnam is willing to make explicit.

Finally, points two and three entail that Putnam cannot (legitimately) use insight into the sociology of mathematics to support an account of axioms which would help him to defend premise $2'$. Even if Putnam could show that axioms serve to clarify mathematical notions by providing implicit definitions of our mathematical vocabulary, and even if he is right that this use of axioms provides the *best* mechanism for fixing the "intended interpretation" of set theory, this still won't give him $2'$. To get $2'$, Putnam would have to show that mathematicians only use axioms to provide *first-order* implicit definitions (point 2); this,

[20] The first two of these points were also made in section 3.2.4. Since each of them is both short and important, it is convenient to repeat them here. See 3.2.4 for a more extensive discussion of these two points.

however, directly conflicts with actual mathematical practice (point 3). Hence, no sociological argument for $2'$ can ultimately be sustained.

How might Putnam respond to this criticism? It seems to me that there are two lines he might take. First, he might deny that his attention to axioms is supposed to be based on *sociological* considerations. In particular, he might claim that, to the extent that there are differing ways of giving background semantics to our axioms, *each of these ways* gives rise to an "intended interpretation" of these axioms. On this line, then, the actual *practice* of mathematicians is more-or-less irrelevant. The mere fact that the axioms of set theory *can* be given a first-order semantics is enough to make Putnam's argument for $2'$ go through.

Note that if Putnam takes this line, he might think that my distinctions between trivial, rich and first-order semantics actually *supports* his overall argument. After all, Putnam showed that any interesting set of axioms has a whole menagerie of first-order "intended interpretations"; my examples of "trivial semantics" show that these axioms have *even more* "intended interpretations." Hence, to the extent that Putnam's argument shows that things are *bad* for the realist, my examples show that things are *even worse.*

There are three things which should be said concerning this bluff response to my criticism of Putnam's argument. First, this is clearly a response which has some appeal for Putnam. At several places Putnam hints that he's "holding back" in using only first-order model theory to find the intended interpretation(s) of set theoretic language. He suggests, for instance, that if we were willing to use intuitionistic model theory, we would find even more "intended interpretations" of set theory. Given this, my examples of trivial semantics should simply provide more grist for Putnam's mill.

Second, this response makes Putnam's overall argument even less plausible than it was before. Whatever worries we might have about countable models of ZFC, it's hard to generate much concern about the fact that *any* model can "satisfy" ZFC in the sense of $\models_{\mathbb{M},\Gamma}$ (or that $\mathbb{S}$ can "satisfy" ZFC in the sense of $\models_s$). If Putnam's argument turns out to be *this* cavalier about the notion of "intended interpretations," then it's hard to take Putnam's argument very seriously.

Third, this response abandons any pretense of giving a genuinely *model-theoretic* argument against realism. Part of the appeal of Putnam's argument stems from his claim that basic theorems of model theory—e.g., the Löwenheim-Skolem theorems, the Shoenfield absoluteness theorem, etc.—entail that set theory has many different "intended interpretations." When filled out in the way we have been discussing, however, the argument actually depends on the more-or-less trivial observation that if you get to reinterpret *anything you want, any way you want*, then you can make *any sentences you want* true under *any circumstances you want.* We didn't need model theory to tell us that.

To avoid this trivialization of his argument, Putnam needs a different strategy for shoring up the "if axioms can't. . . " defense of premise $2'$. In particular, he needs a strategy which lets him argue that first-order semantics are *canonical* with respect to picking out "intended interpretations." Such a strategy will involve (at least) two things: 1.) an argument that first-order semantics *always* generates intended interpretations for a set of axioms (even when those axioms seem to presuppose a "richer" semantics) and

2.) a proof that this argument doesn't ramify so as to allow trivial forms of semantics to generate still more intended interpretations.

There are two things to note about this second strategy. First, it isn't enough for Putnam to simply argue that certain axioms *can be* reinterpreted so as to take on first-order semantics. After all, these same axioms *can be* reinterpreted so as to take on various kinds of trivial semantics (e.g., $\models_{\mathbb{M},\Gamma}$ or $\models_s$). Instead, Putnam needs to show that there is something special about first-order semantics: something which makes first-order model theory—and *only* first-order model theory—semantically normative.

Second, something like this second strategy is necessary *whether or not* Putnam relies on the "if axioms can't ..." defense of premise $2'$. Whatever Putnam's defense of $2'$ happens to be, if it leads to the conclusion that trivial semantics generate intended interpretations, then Putnam's overall argument is in trouble (for the reasons discussed above). Thus, *any* argument for $2'$ will have to satisfy the two conditions mentioned at the end of the penultimate paragraph. It will have to show that first-order reinterpretations of sentences which seem to presuppose richer background semantics really count as *intended* interpretations, and it will have to explain why this claim doesn't extend to make trivial reinterpretations count as intended.

Now, as far as I know, Putnam has never seriously attempted to give this kind of argument. As a result, his model-theoretic argument against realism is incomplete. Further, this is not an incompleteness which can easily be remedied: as far as I can see, there simply *is* no good reason for thinking that first-order semantics has some special normativity vis-a-vis the interpretation of set-theoretic axioms. More significantly, I think it is impossible to give an argument for such normativity that will square with Putnam's *own* theoretical commitments. To understand this impossibility, we need to move beyond the "if axioms cannot..." defense of premise $2'$ to examine Putnam's third (and final) defense of that premise.

### 4.4.3 Just more theory I

The "just more theory" argument is Putnam's most famous (and boldest) defense of premise $2'$. It has been deployed most explicitly—and been subject to the most analysis—in cases where *causality* is supposed to pick out the interpretation of a piece of language. Nevertheless, the structure of the argument is quite general and applies to any realistic account of the nature of reference.[21]

The substance of the "just more theory" argument lies in the observation that the phrase "theoretical constraints" is broad enough to cover *philosophy* as well as mathematics and natural science. In particular, any philosophical account of the way set theory gets its "intended interpretation" can itself be viewed as just one more theoretical constraint. Hence, no such philosophical account can be adequate: since Putnam can find an assortment of models which satisfy *both* our original theoretical constraints *and* our philosophical semantics, the philosophical semantics cannot force our language to take on a unique "intended interpretation."

[21] For commentary on the "just more theory" argument, see [7], [11], [21], [22], [31], and [60]. For some recent revisionary interpretations of the argument, see [1] and [12].

To illustrate this point, consider the following theory (or schematic of a theory) concerning the "intended interpretation" of set theory:

**HYP** In order to be an intended interpretation for the language of set theory, a structure must satisfy conditions $C_1$,. . . , $C_n$.

There are two things we should notice about this schematic. First, it is relatively easy to come up with conditions $C_1$,. . . , $C_n$ which restrict—and restrict in a fairly plausible fashion—the structures which count as intended interpretations for the language of set theory. At the simplest level, we can insist that such structures be well-founded or transitive.[22] More boldly, we can insist that such structures satisfy second-order ZFC or that they correctly interpret some primitive notation for power-sets (e.g., up to isomorphism). Finally, and most audaciously, we can insist that any intended interpretation for the language of set theory include all and only sets in its domain and interpret "∈" as referring to real, set-theoretic membership.

If Putnam's argument is correct, however, none of these conditions help the realist. If **HYP** is a good theory concerning the interpretation of set-theoretic language, then **HYP** should be added to our overall theory of the world. In Putnam's terms, **HYP** should be added to our "theoretical and operational constraints." But, since premise $1'$ already takes care of theoretical and operational constraints, this addition does not effect Putnam's overall argument. Because the phrase "theoretical and operational constraints" is flexible enough to *commandeer* rival mechanisms for fixing the intended interpretation of set theory, no such mechanisms can possibly threaten Putnam's argument.

It is useful, at this point, to sketch a fairly standard response to this move of Putnam's.[23] The response has two parts. First, we draw a distinction between *describing* the features of a model which make it an "intended interpretation" and simply *adding sentences* which a model must satisfy in order to be an intended interpretation. Put more perspicuously: we distinguish between *changing* the semantics under which a collection of axioms is interpreted—e.g., by restricting the class of models for our language and/or strengthening the notion of "satisfaction" which ties sentences to models—and simply adding new sentences which get interpreted under the same old semantics.

Second, we note that the proponent of conditions like those discussed above—i.e., in the examples of ways to fill out $C_1, \ldots, C_n$—typically uses those conditions to *describe* the features which make a model intended. In other words, she is using them to explain *which semantics* she intends her axioms to be interpreted under, not just listing additional sentences which are to be interpreted under the (first-order) semantics of Putnam's choosing. If, for instance, condition $C_1$ says that a model should be well-founded, then the proponent of

[22] Note that, although these are fairly weak conditions, they are strong enough to eliminate many of the "intended interpretations" proposed by Putnam (e.g., the one induced by Theorem 1). This highlights the point, made in footnote 1, that Putnam's model-theoretic argument involves a *radical* form of semantic indeterminacy: even weak restrictions on the structures which serve as "interpretations" for our language will eliminate the kinds of indeterminacy which this argument purports to uncover.

[23] This response has its home in the literature concerning the relationship between Putnam's argument and the causal theory of reference. Its structure is quite general however. See [11] or [31].

**HYP** will demand that candidate models *actually be* well-founded. She will not consider it enough for these models to satisfy a first-order axiom of the form: $\neg\exists\, x\,(x \textit{ is an infinite descending chain of ordinals})$.

Now, as noted above, this kind of objection to the "just more theory" argument is quite well known. It is not surprising, therefore, that Putnam has a well worked-out response. Basically, Putnam claims that this objection *begs the question* against the model-theoretic argument. The argument, after all, involves the question of whether mathematical language has a unique "intended interpretation." Hence, realists cannot *assume* that mathematical language has such an interpretation when they try to answer Putnam. In particular, they cannot assume that words and phrases like "well-founded," "transitive," and "complete power-set" have determinate interpretations when they set out to "describe" their notion of intended models (or to "explain" the semantics under which they intend their axioms to be interpreted). To paraphrase a passage of Putnam's:[24]

> Here the philosopher is ignoring his own epistemological position. He is philosophizing as if naive realism were true of him (or, equivalently, as if he and he alone were in an *absolute* relation to the world). What *he* calls 'transitivity' really is transitivity, and *of course* there is a fixed, somehow singled-out, correspondence between the word and one definite property in *his* case. Or so he assumes. But how this can be so was just the question at issue. ([44] xi)

In Putnam's view, then, as long as the determinacy of mathematical language is still in question, realists cannot—on pain of begging the question—use this language to specify the intended interpretation of their axioms. As a result, they cannot use claims like **HYP** to *describe* the intended interpretation of set theoretic language; instead, they have to view **HYP** as a mere addition to their "theoretical constraints." But this amounts, in Putnam's words, to the addition of "just more theory."

To summarize: Putnam's "just more theory" defense of premise $2'$ comes in two stages. First, Putnam considers realistic theories of interpretation and argues that these theories can be regarded as additional "theoretical constraints." As such, they do they do not add anything new to the "theoretical and operational constraints" of premise $1'$. Second, Putnam considers the claim that realistic theories *describe* the ways interpretations are fixed (or, equivalently, describe the kinds of semantics under which our axioms are to be interpreted). Since this claim assumes that such description is ultimately possible—i.e., that we can use mathematical language to describe, in a determinate fashion, the intended interpretation of set theory—Putnam dismisses it as question-begging.[25]

[24] My paraphrase here consists solely of tailoring a passage of Putnam's to make it fit the mathematical case. In particular, I have replaced the term "causality" with "transitivity" in two places, and replaced the phrase "one definite relation" with the phrase "one definite property." Clearly these changes do not affect the argumentative structure of the passage.

[25] The interpretation of the "just more theory" argument which I have given is fairly standard (see [11], [31], and [60]). In [1] and [12], Anderson and Douven challenge this interpretation. Without going into too many details, I find these revisionary interpretations of Putnam's argument uncompelling. For one thing, I don't think they fit Putnam's texts as well as the standard interpretation (especially Putnam's comments in the introduction to [44] and section 8 of [43]. For another, I think they both make Putnam's argument *less* interesting than it was on the standard interpretation. Hence, even if these revisionary interpretations *did* fit the texts, we would still want to focus our attention on the standard interpretation.

### 4.4.4 Just more theory II

In this section, I give three responses to Putnam's "just more theory" argument. First, I show that this argument is incompatible with some of Putnam's other theoretical commitments; hence, whatever merits it may have in its own right, it is not an argument which is available to Putnam. Second, I show that an analogous argument can be made in situations where it *clearly* doesn't work, and I argue that, by parity of reasoning, Putnam's own argument doesn't work either. Finally, I give a straightforward explanation as to why Putnam's argument doesn't work: in essence, I show that Putnam's charge of question begging rests on a (fairly trivial) logical mistake. At the end of the section, then, I conclude that Putnam's "just more theory" argument provides no support for premise $2'$.

To begin, note that Putnam's charge of question begging in the last section both introduces and rests upon a certain assumption: that it is illegitimate to use semantically indeterminate language to describe the intended interpretation of set theory. After all, the problem with realists' attempts to use conditions like $C_1, \ldots, C_n$ to describe the interpretation of set theory was that they could not show that these conditions had determinate interpretations (and that simply *assuming* that they had determinate interpretations would beg the question). For this objection to have force, however, we must first think that there is something *wrong* with using semantically *in*determinate language to describe the intended interpretation of set theory. Only against the backdrop of this assumption would the realist even *need* to claim that $C_1, \ldots, C_n$ are semantically determinate.

Next, we need to recall three technical facts about first-order logic. First, this logic lacks the resources to capture the notions of finitude and/or recursion. On the finitude side, there is no collection of formulas $\Gamma$ which will characterize exactly the finite models (or finite subsets of models). On the recursion side, there is no collection of formulas $\Gamma$ which lets us capture the notion of a *recursive* definition.[26] Second, the notion of finitude is needed to characterize the syntax of first-order logic: the sentences of first-order logic can have arbitrarily *finite* length, but they cannot be infinite.[27] Third, the standard definition of first-order satisfaction is recursive: it starts with a definition of satisfaction for atomic formulas and then supplies recursion clauses—one for each connective and quantifier in our language—to extend this definition to arbitrary formulas.

Combining these technical facts with the assumption from the antepenultimate paragraph, it should be obvious that Putnam has gotten himself into hot water. The following four claims are on the table:

1. The notions of finitude and recursion are needed to describe first-order model-theory.
2. First-order model theory cannot capture the notions of finitude and recursion.

[26] For details on the problem with the notion of finitude, see p. 57. This problem extends to recursion, because recursive definitions try to capture a finite—but arbitrarily finite—number of steps in a single definition.

[27] It is important, here, to notice that first-order ZFC makes explicit use of this convention. Insofar as two of its axioms—comprehension and replacement—are really axiom schemas (see fn. 56 in ch. 3), these axioms cannot even be formulated without a firm grasp on the notion of finitude.

3. It is illegitimate to use semantically indeterminate notions to describe "intended interpretations."
4. *Only* those notions which can captured by first-order model theory are semantically determinate.

Here, claims 1 and 2 are just repetitions of the technical facts discussed in the last paragraph, and claim 3 is the assumption on which Putnam's charge of "begging the question" was seen to rest. Claim 4 is a version of Putnam's own premise $2'$. It is needed to ensure that realists do not use axioms with a rich semantics to pin down the intended interpretation of set theory, and it was discussed in a fair bit detail near the end of section 4.4.2.

Together, these four claims cause serious problems for Putnam. From claims 3 and 4, it follows that only notions which can be captured by first-order model theory can be used to describe the "intended interpretation" of set theory. Combining this with claim 2, we get that the notions of finitude and recursion cannot be used to describe the intended interpretation of set theory. Together with claim 1, this entails that first-order model theory cannot be used to describe the intended interpretation of set theory. Since this is precisely what Putnam's own argument requires us to do (cf. 4.4.2), Putnam's overall position reduces to a contradiction.

This, then, is what I call the "stability" objection to Putnam's "just more theory" argument. The argument rests on the assumption that we cannot use semantically indeterminate language to describe "intended interpretations." But, by Putnam's *own standards,* the notions needed to formulate first-order model theory turn out to be semantically indeterminate. Hence, since Putnam's own technique for obtaining intended interpretations of set theory involves applying first-order model theory to the axioms of ZFC, his overall position is internally inconsistent.

Leaving this stability objection aside, we should note that portions of Putnam's "just more theory" argument can be viewed as structurally analogous to arguments concerning other puzzles I have examined in this dissertation. Consider, for instance, the "paradox of the three men" from section 3.1.3. Suppose that someone proposes this paradox as a serious challenge to the determinacy of ordinary talk about the natural numbers, and suppose that we respond to this challenge by solving the paradox in the manner sketched in 3.1.3. Why, in this case, couldn't the proponent of the paradox simply follow Putnam and claim that our solution "begs the question"? After all, the paradox is supposed to challenge the determinacy of ordinary talk about numbers, and we assume the determinacy of such talk for the purpose of giving our solution.

A similar problem can be raised concerning the supermodel construction from section 3.2.3. Suppose that someone suggests that the supermodel should count as an "intended interpretation" for the language of set theory. If we respond that this model has obvious defects—e.g., it contains only one thing and it fails to respect bivalence—our interlocutor could claim that this response is "just more theory." What's more, he can easily show that the supermodel takes care of such theory. After all, the supermodel trivially (super)satisfies all sentences of the following forms: $\exists x \exists y(\neg x = y)$, $\exists x \exists y \exists z(\neg x = y \wedge \neg y = z \wedge \neg x = z)$, $\phi \vee \neg \phi$, $\neg(\phi \wedge \neg \phi)$, etc.

At this point, we might object that our interlocutor has failed to understand our initial response. We weren't interested in whether $\mathbb{S} \models_s \exists x \exists y\, (\neg x = y)$; we were interested in the fact that $\mathbb{S}$ contains only *one thing.* Similarly, we weren't interested in whether $\mathbb{S} \models_s \phi \vee \neg\phi$; we were interested the fact that $\mathbb{S}$ fails to respect the *pattern,*

$$\mathbb{S} \models_s \phi \Longleftrightarrow \mathbb{S} \not\models_s \neg\phi.$$

Here again, though, our interlocutor can simply follow Putnam. Since our objection assumes that we can talk about sets in a determinate manner—e.g., those sets needed to formulate notions like "one-element domain"—it seems to beg the question against the advocate of the supermodel. Why, then, shouldn't our objection be summarily dismissed?

At this point, therefore, we have two arguments that are structurally analogous to the "just more theory" portion of Putnam's model-theoretic argument. Clearly, neither of them is very effective. The "paradox of the three men" is not a real paradox; the solution given in 3.1.3 is perfectly adequate; and we do not beg any questions in giving this solution. Similarly, the supermodel is not an "intended interpretation" for the language of set theory; the fact that it has a one-element domain is part of *why* it is not an intended interpretation; the fact that it supersatisfies $\exists x \exists y(\neg x = y)$ is irrelevant; and we don't beg any questions in saying all of this. Hence, given the obvious structural similarities between these two failed arguments and Putnam's "just more theory" argument, I think we have ample grounds for rejecting this latter argument (if only for reasons of parity).

Of course, it would be nice if we didn't have to rely on parity—or on theoretical conflicts with *Putnam's* other views—to get a response to the "just more theory" argument. Fortunately, it's relatively easy to explain why Putnam's charge of begging the question misses the mark. In evaluating Putnam's model-theoretic argument we are interested in the following question: does Putnam's model theory entail that mathematical language is semantically indeterminate? To answer this question, we have to evaluate a conditional of the following form:

$$\text{Putnam's Model Theory } \Longrightarrow \text{Semantic Indeterminacy.}$$

Now, as a general rule, conditionals of the form $\mathbf{P} \Longrightarrow \neg\mathbf{Q}$ are evaluated by asking whether we can accept both $\mathbf{P}$ and $\mathbf{Q}$. If we can, then the conditional has to be rejected.

Turning to the case at hand, therefore, we ask whether we can accept both Putnam's model theory and the semantic determinacy of mathematical language. That is, we (tentatively and hypothetically) accept these things, and we then check to see whether the resulting combination leads to a contradiction. In this context, therefore, our hypothetical acceptance of semantic determinacy does not constitute an instance of "begging the question." It's just part of the way we go about evaluating conditionals.[28]

[28] Similar comments apply to the two examples discussed a moment ago. It's precisely because the paradox of the three men is a puzzle in arithmetic that we should feel free to use arithmetic when solving it. Similarly, it's because the supermodel puzzle involves the relationship between model theory and set theory that we use these two disciplines in solving *that* puzzle.

To make this point somewhat more perspicuous, let me give an example which does not involve any mathematics (nor, indeed, any questions about semantic determinacy). Suppose that two philosophers—John and Alvin—are arguing about the existence of God. John presents a version of the problem of evil to show that God does not exist, and Alvin tries to defuse this problem by proving that none of John's evils are really incompatible with the existence of God. In particular, Alvin sketches out a hypothetical situation in which both God and John's evils exist at the same time.

In this example, Alvin has not "begged the question" against John. Even though the overall debate is about the existence of God, it is perfectly legitimate for Alvin to evaluate John's argument by constructing a hypothetical situation in which God is assumed to exist. Although this argument can't show that God *does* exist—so it can't show that Alvin's theodicy is *true*—it can show that John's argument against the existence of God fails.

So too, then, in the case of Putnam's model-theoretic argument. Even though the overall debate between Putnam and the realist concerns the determinacy of mathematical language, it is perfectly legitimate for the realist to presuppose such determinacy in the course of evaluating the validity of Putnam's argument. Just as Alvin's argument begs no questions against John, so too the realist's argument begs no questions against Putnam. To think otherwise, is to misunderstand the logic of ordinary conditionals.

Before closing out this section, a final comment on this "evaluation of conditionals" point is in order. None of my discussion of this point depends on any controversial assumptions about the status of the larger dialectic between Putnam and the realist. The point does not depend on the assumption that we have an adequate account of mathematical reference. Nor does it depend on fancy considerations concerning the burden of proof in realist/anti-realist debates. Indeed, the point is even compatible with the assumption that reference to mathematical objects is indeterminate—i.e., with the assumption that Putnam is *right* on the deeper philosophical issue. The point simply shows that one particular aspect of Putnam's model-theoretic argument—namely, his accusation that realists "beg the question" in their response to his "just more theory" maneuver—rests on a trivial misapplication of the logic of conditionals.

To conclude, then, Putnam's "just more theory" argument fails. There is a basic difference between *describing* the features which make a model "intended" and simply *adding* new sentences for a model to satisfy. Put another way, there is a difference between *changing* the semantics under which certain axioms get interpreted and adding new sentences to be interpreted under the same *old* semantics. Further, Putnam's charges of "begging the question" don't allow him to undercut—or to slur over—this distinction. At best, these charges rest on a misapplication of the logic of ordinary conditionals. At worst, they introduce assumptions which get Putnam into even more trouble than he was in before—e.g., the stability objection.

## 4.5 Some Connections

In this section, I want to draw some connections between the philosophical problems discussed in 4.4 and the mathematical problems discussed in 4.3. In particular, I will show that both kinds of problems involve a certain dialectical assumption that Putnam tends to make, and I will argue that realists have no reason to grant Putnam this assumption. In light of this, I will argue, once again, that there is very little reason for realists to worry about Putnam's model-theoretic argument.

I begin with a passage originally discussed in section 4.4.3. Here, Putnam is considering the claim that realists use mathematical language to *describe* the intended interpretation of set theory, and he argues that this claim "begs the question" against the model-theoretic argument. Putnam complains: "Here the [realist] is ignoring his own epistemological position. He is philosophizing as if naive realism were true of him. . . as if he and he alone were in an *absolute* relation to the world." Now, leave aside the question of whether there is anything *wrong* with such naive philosophizing—a question answered in the negative in the last section. Instead, ask whether Putnam himself can give the model-theoretic argument without engaging in naive philosophizing of his own.

To see the worry here, recall the overall structure of Putnam's argument. Putnam begins by considering a theory which includes "all theoretical and operational constraints." In principle, this theory includes *everything* we might ever need to say; indeed, at one point, Putnam suggests that it includes *all true sentences* ([43] 18). Once this theory is in hand, Putnam argues that it has many different models; further, the fact that these models satisfy "all theoretical and operational constraints" ensures that each of them counts as an "intended interpretation" of our theory.

Why, though, does Putnam think that we can discuss these models in the first place? Insofar as the models provide the (or, perhaps, a) semantics for a particular theory, it's not obvious that they can be discussed from *within* that theory. Hence, Putnam's own argument seems to assume that we can, at least to a certain extent, "stand back" from our theory in order to discuss that theory's semantics. But, given that the theory in question is supposed to constitute our *entire* theory of the world—to satisfy *all* theoretical and operational constraints and to include *all* true sentences—it's not clear how this "standing back" is supposed to work.

This, then, is where Putnam's own "naive realism" comes into play. Whatever problems our theory is supposed to have, these problems vanish when Putnam steps back from this theory to engage in semantics. Putnam allows himself to refer, naively and absolutely, to both the theory in question and to a collection of models for this theory. He also allows himself the full apparatus of classical model theory—e.g., everything needed to make sense of the notion of satisfaction—in order to talk about a theory's being true or false "on a certain interpretation."

Of course, Putnam thinks that there are many different interpretations which make our theory come out true (that, after all, is supposed to be the *point* of the model-theoretic argument). But, for each particular "interpretation," Putnam spells out the relationship between that interpretation and the "truth" of our theory in a thoroughly realistic manner. To paraphrase: Putnam philosophizes as though he and he alone were in an *absolute* relation to the world of models—what Putnam calls "models" really are models, what Putnam calls "satisfaction" really is satisfaction, and *of course* there is a fixed, somehow singled-out, correspondence between the language of model theory and the world of models in *his* case.

As we saw in 4.4.3, however, this is precisely the standpoint which Putnam wants to deny to the realist. When the realist tries to "stand back" from set theory in order to talk about this theory's intended interpretation—to specify, for instance, that the interpretation must *really* involve sets, must *really* be well-founded, must *really* satisfy second-order ZFC, or must *really* include all ordinals—Putnam accuses him of "begging the question." Though Putnam's own model-theoretic talk can be viewed as talk *about* set theory, the realist's talk must be viewed as talk *within* set-theory.

This asymmetry leaves Putnam with a dilemma. On the one hand, Putnam can allow people to "stand back" from their theories in order to discuss these theories' semantics. In this case, Putnam himself can stand back from set theory in order to cook up non-standard models for this theory, but the realist can also stand back to explain why these models do not constitute "intended interpretations." Just as Putnam uses absolute model-theoretic notions—like "model," "interpretation," "satisfaction," etc.—to show that set theory has many models, so the realist can use notions like "well-founded," "uncountable" and even "set" to explain why most of these models are unintended.

On the other hand, Putnam can stick to his guns and insist that people can't "stand back" from their theories in order to discuss these theories' semantics. In this case, the realist would be barred from arguing that Putnam's unusual models are "unintended" because they fail to satisfy certain constraints which the realist has (naively) laid down. But, Putnam himself will be barred from claiming that any such unusual models even exist! Hence, he will be barred from posing the question which lies at the heart of the model-theoretic argument: why don't these (strange, unusual, and ugly) models constitute "intended interpretations" for the language of set theory?

This, then, is Putnam's dilemma. Two comments about it are in order. First, the dilemma clearly amounts to a less-technical reworking of some of the ideas behind the "stability objection" of section 4.4.4. There, I tried to how the asymmetry in Putnam's argument leads to a fairly specific inconsistency in Putnam's position. In particular, Putnam's argument assumes that we can use the notions of finitude and recursion to pick out "intended interpretations" for set theory; but, if Putnam's arguments are correct, then these are exactly the kinds of notions that realists *cannot* cannot use to pick out intended interpretations.[29] Hence, it is only by assuming a stark asymmetry between his own position and that of his (realist) opponents that Putnam can make his overall argument look consistent.

Second, this dilemma bears an evident similarity to the dilemma discussed at the end of section 4.3.1. There, Putnam wanted to limit his opponents to a certain collection of set-theoretical axioms, while he himself

[29] i.e., since 1.) these notions can't be captured by first-order model theory, 2.) only notions that can be captured by first-oder model theory are determinate, and 3.) only determinate notions can be used to specify intended interpretations.

used somewhat stronger axioms to prove his key theorem. Here, Putnam wants to limit his opponents to working *within* a particular theory, while he himself steps outside this theory to talk about its semantics. In both cases, then, Putnam's argument depends on allowing himself just a little more material than he allows the realists against whom he is arguing.

This, then, suggests that there is a unified response which realists can give to Putnam's two arguments—i.e., to his arguments for premises $1'$ and $2'$. In both cases, realists can simply insist that Putnam use the same—and only the same—resources in formulating his arguments which he allows the realist to use in responding to them. Applying this principle at the technical level undercuts Putnam's defense of premise $1'$. Applying this principle at the philosophical level undercuts his defense of premise $2'$. With *both* of these premises so disabled, Putnam's model-theoretic argument poses little threat to traditional realism.[30]

## 4.6 Conclusion

At the end of the day, what should we make of Putnam's model-theoretic argument? On the one hand, the argument has two real virtues. It brings out clearly the degree to which Skolem's Paradox—in either it's classical form or in the modified form discussed here—is ultimately a problem in *semantics.* As such, it fits in well with the main theme of section 1.2.2. The argument also frames the relevant semantic problem—the problem of determinate reference for mathematical language—in a way which has captured the attention of a great many contemporary philosophers. For this, we should all be grateful.

However, as an outright argument for semantic anti-realism, Putnam's model-theoretic argument leaves a lot to be desired. As a *proof* of anti-realism, the argument fails for the, essentially technical, reasons discussed in 4.3. Even as a suggestive philosophical argument it doesn't really work. To be sure, if the only way of fixing the reference of mathematical language is through the sort of massive first-order "implicit definition" involved in premise $1'$, then the argument raises threatening possibilities. But there is *no reason* for thinking that this *is* the only way of fixing reference. Putnam's own arguments for this claim are, as I have shown in 4.4–4.5, singularly unpersuasive. Further, I think *any* arguments for this claim are liable to run aground on problems like the "stability objection" discussed in 4.4.4.

When all is said and done, therefore, the question of reference to mathematical objects remains. It's a hard question, and realists will have a hard time—for the reasons discussed in section 3.1—giving a plausible answer to it (i.e., giving a substantive account as to just how such reference is supposed to work). But,

[30] I have placed this argument at the end of this chapter, because I think it provides a useful way of tying together the main themes of sections 4.3 and 4.4. I do not, however, think that this kind of "dialectical" argument provides the *best* response to Putnam's arguments. The best response, I think, is to argue directly against Putnam's non-standard models for set theory (by noting, e.g., that their non-transitivity or their inability to correctly interpret the definition of power sets makes them unsuitable as "intended interpretations" of set theory). If Putnam claims that this argument begs the question—as he did, for instance, in section 4.4.3—this claim should be defused by appealing to the evaluation-of-conditionals argument from section 4.4.4. By responding in this way, we defuse Putnam's model-theoretic argument without appealing to the details of our overall dialectical situation.

however this question is ultimately answered, I think I have shown that Skolem's Paradox—whether in its classical form or in the form of Putnam's model-theoretic argument—will play no real role in this answer. That is enough for this thesis.

# Bibliography

[1] David Anderson. What is the model-theoretic argument. *The Journal of Philosophy*, 93:311–22, 1993.

[2] Paul Benacerraf. Mathematical truth. In *Philosophy of Mathematics* [5], pages 403–420.

[3] Paul Benacerraf. What the numbers could not be. In *Philosophy of Mathematics* [5], pages 272–294.

[4] Paul Benacerraf. Skolem and the skeptic. *Proceedings of the Aristotelian Society*, 59:85–115, 1985.

[5] Paul Benacerraf and Hilary Putnam. *Philosophy of Mathematics*. Cambridge University Press, Cambridge, 1983.

[6] George Boolos. The iterative conception of set. *The Journal of Philosophy*, 68:215–230, 1971.

[7] Anthony Brueckner. Putnam's model-theoretic argument against metaphysical realism. *Analysis*, 44:134–40, 1984.

[8] Cesare Burali-Forti. A question on transfinite numbers. In van Heijenoort [61], pages 104–112.

[9] Rudolf Carnap. *Einführung in die Symbolische Logik*. Springer, Vienna, 1954.

[10] Dauben. *Georg Cantor*. Harvard, Cambridge, 1979.

[11] Michael Devitt. *Realism & Truth*. Princeton University Press, Princeton, 1984.

[12] Igor Douven. Putnam's model-theoretic argument reconstructed. *The Journal of Philosophy*, 96:479–90, 1999.

[13] F. Drake. *Set Theory: An Introduction to Large Cardinals*. North Holland, Amsterdam, 1974.

[14] Michael Dummett. The reality of the past. *Proceedings of the Aristotelian Society*, 69:239–258, 1969.

[15] John Echemendy. *The Concept of Logical Consequence*. Harvard, Cambridge, 1990.

[16] Hartry Field. Fictionalism, epistemology and modality. In *Realism, Mathematics and Modality*, pages 1–52. Blackwell, Oxford, 1989.

[17] Hartry Field. Are our logical and mathematical concepts highly indeterminate. *Midwest Studies in Philosophy*, 19:391–429, 1994.

[18] Arthur Fine. Quantification over the real numbers. *Philosophical Studies*, 19:27–31, 1968.

[19] Alexander George, editor. *Reflections of Chomsky.* Blackwell, Cambridge, 1989.

[20] Michael Hallett. *Cantorian Set Theory and Limitation of Size.* Oxford, Oxford, 1984.

[21] Craig Hansen. Putnam's indeterminacy argument: The Skolimization of absolutely everything. *Philosophical Studies*, 51:77–99, 1987.

[22] Mark Heller. Putnam, reference and realism. *Midwest Studies in Philosophy*, 12:113–28, 1988.

[23] Geoffrey Hellman. *Mathematics without Numbers.* Clarendon Press, Oxford, 1989.

[24] Geoffrey Hellman. Structuralism without structures. *Philosophy of Mathematics*, 4:100–123, 1996.

[25] Wilfrid Hodges. Truth in a structure. *Proceedings of the Aristotelian Society*, 86:135–151, 1985–86.

[26] Thomas Jech. *Set Theory.* Academic Press, San Diego, 1978.

[27] Philip Kitcher. The plight of the platonist. *Noûs*, 12:119–136, 1978.

[28] Virginia Klenk. Intended models and the Löwenheim-Skolem Theorem. *The Journal of Philosophical Logic*, 5:475–489, 1976.

[29] Julius Konig. On the foundations of set theory and the continuum problem. In van Heijenoort [61], pages 143–149.

[30] Kenneth Kunnen. *Set Theory.* North-Holland, Amsterdam, 1980.

[31] David Lewis. Putnam's paradox. *Australasian Journal of Philosophy*, 62:221–236, 1984.

[32] Penelope Maddy. *Realism in Mathematics.* Oxford, New York, 1990.

[33] Van McGee. How do we learn mathematical language. *The Philosophical Review*, 106:35–68, 1997.

[34] Christopher Menzel. Choice again. *Philosophical Studies*, 49:37–61, 1986.

[35] Moore. *Zermelo's Axiom of Choice.* Springer-Verlag, New York, 1982.

[36] Yiannis Moschovakis. The formal language of recursion. *The Journal of Symbolic Logic*, 54:1216–1252, 1989.

[37] Charles Parsons. The structuralist view of mathematical objects. *Synthese*, 843:303–346, 1990.

[38] Stephen Pollard. On the itterative explanation of the paradoxes. *Philosophical Studies*, 66:285–296, 1992.

[39] Hilary Putnam. The thesis that mathematics is logic. In *Mathematics, Matter and Method*, pages 12–42. Cambridge, New York, 1975.

[40] Hilary Putnam. *Meaning and the Moral Sciences.* Routledge, New York, 1978.

[41] Hilary Putnam. Realism and reason. In *Meaning and the Moral Sciences* [40], pages 123–138.

[42] Hilary Putnam. *Reason, Truth and History.* Cambridge University Press, New York, 1981.

[43] Hilary Putnam. Models and reality. In *Realism and Reason* [44], pages 1–25.

[44] Hilary Putnam. *Realism and Reason.* Cambridge UP, Cambridge, 1983.

[45] Hilary Putnam. Model theory and the 'factuality' of semantics. In George [19], pages 213–231.

[46] Willard Van Orman Quine. New foundations for mathematical logic. *American Mathematical Monthly*, 44:70–80, 1937.

[47] Michael Resnik. On skolem's paradox. *The Journal of Philosophy*, 63:425–438, 1966.

[48] Michael Resnik. *Frege and the Philosphy of Mathematics.* Cornell, Ithica, 1980.

[49] Michael Resnik. *Mathematics as a Science of Patterns.* Oxford, New York, 1997.

[50] Jules Richard. The principles of mathematics and the problem of sets. In van Heijenoort [61], pages 142–144.

[51] Bertrand Russell. *The Principles of Mathematics.* Cambridge, London, 1903.

[52] Betrand Russell. Mathematical logic as based on the theory of types. *American Journal of Mathematics*, 30:222–262, 1908.

[53] Stewart Shapiro. *Foundations without Foundationalism.* Clarendon, Oxford, 1991.

[54] Thoralf Skolem. *Selected Works in Logic.* Uiversitetsforlaget, Oslo, 1970.

[55] Alfred Tarski. Foundations of the calculus of systems. In *Logic, Semantics, and Metamathematics* [56], pages 342–383.

[56] Alfred Tarski. *Logic, Semantics, Metamathematics.* Hackett, Indianapolis, 2nd edition, 1983.

[57] Alfred Tarski. On the concept of logical consequence. In *Logic, Semantics, and Metamathematics* [56], pages 409–20.

[58] William Thomas. Platonism and the skolem paradox. *Analysis*, 28:193–196, 1968.

[59] William Thomas. On behalf of the skolemite. *Analysis*, 31:177–186, 1971.

[60] James Van Cleve. Semantic supervenience and referential indeterminacy. *The Journal of Philosophy*, 89:344–361, 1992.

[61] Jean van Heijenoort, editor. *From Frege to Gödel*. Harvard, Cambridge, Mass., 1967.

[62] Hao Wang. The formalization of mathematics. *The Journal of Symbolic Logic*, 19:247, 1954.

[63] Hao Wang. *A Survey of Mathematical Logic*. North-Holland, Amsterdam, 1964.

[64] Hao Wang. The concept of set. In *Philosophy of Mathematics* [5], pages 530–570.

[65] Crispin Wright. Skolem and the skeptic. *Proceedings of the Aristotelian Society*, 59:116–137, 1985.

[66] Ernst Zermelo. Sur les ensembles finis et le principe de l'induction complete. *Acta Mathematica*, 32:185–193, 1909.

[67] Ernst Zermelo. Über grenzzahlen und megenbereiche: neue untersuchungen über die grundlagen der mengenlehre. *Fundamenta Mathematicae*, 16:29–47, 1930.

[68] Ernst Zermelo. Investigations in the foundations of set theory I. In van Heijenoort [61], pages 199–215.