2

Extra-Worldly Semantics

2.1 Introduction

In this chapter, I'll consider semantic theories that take the form of what I'll call "extra-worldly semantics." In the paradigmatic case, such a semantic theory will be one in which we think of the meaning of a sentence ϕ in terms of the set of completely determinate ways for the world to be—the set of "possible worlds"—such that each element of that set is a way for the world to be such that ϕ is true.[1] As a variant of worldly semantics, an extra-worldly semantics requires us to try to comprehend our knowledge of meaning sentences and predicates as assymetrically dependent on our knowledge worldly entities and their relations. Specifically, an extra-worldly semantics thinks of this knowledge as knowledge of possible worlds, objects contained within them, set-theoretic relations between possible worlds and the objects contained within them. I will argue that we cannot comprehend our knowledge of meaning in this way. The core problem is that the extra-worldly knowledge to which an extra-worldly semantics appeals is only intellegible as *dependent on* our knowledge of propositions and properties, but this knowledge of propositions and properties, on an extra-wordly semantics, is *understood in terms of* our knowledge of terms of sets of worlds and functions from worlds to extensions. I will go on to argue in the next chapter that this

[1] Or, equivalently as a function from the total set of worlds to the value *true* or *false*. Informational dynamic semantic theories, which I'll discuss briefly in Chapter 4, are also variants of extra-worldly semantics.

knowledge of propositions and properties is itself dependent on our knowledge of the correct use of sentences and predicates, but it suffices for my purposes in this chapter to show that knowledge of propositions and properties cannot be understood in terms of knowledge of possible worlds and the objects contained within them. In the form of the phrase provided in the previous chapter of what is problematic about the Myth of the Given, an extra-worldly semantics requires the availability for extra-worldly knowledge to subjects whose getting this extra-worldly knowledge does not draw on the capacities required for this exra-worldly knowledge.

2.2 The Extra-Worldly Meaning of Predicates

In the previous chapter (§1.2), I introduced a toy language consisting in sentences like "*a* is gray," "*b* is darker than *c*," "It's not the case that *c* is white," "*c* is gray or *b* is black," and so on. Our task is now to articulate a semantic theory for this language that is able to explain the behavior that speakers of it exhibit in virtue of grasping the meanings of the expressions that belong to it. More concretely, our task is to construct a function $[\![\cdot]\!]$ that maps each sentence ϕ of this toy language to its *semantic value* $[\![\phi]\!]$, the element of our theory that is meant to serve as a model of what speakers of this language know in knowing the meaning of ϕ and to do so in such a way that the semantic value of a complex sentence is determined by the semantic values of its parts and the way those parts are put together. The formal framework in which this task is usually undertaken is what I am calling an "extra-worldly" semantic framework, or what is more commonly called a "possible worlds semantics," and that is that is the formal framework that we'll consider in the present chapter.

There are several motivations for an extra-worldly semantics. Perhaps the main motivation is that it functions as a single framework in which we can provide

set-theoretic characterizations of various phenomena of concern to the study of linguistic meaning: semantic relations such as synonymy, entailment, and incompatibility, logical operations such as conjunction, disjunction, and negation, modal notions such as possibility, necessity, permission, and obligation (Kripke, Kratzer), counterfactuals (Lewis), pragmatic phenomena such as the effect of making an assertion on a state of inquiry (Stalnaker). All of this, however, hangs on the thought that simple content words such as the basic predicates of a language can be reasonably assigned semantic values in a possible worlds semantics. It is the assignment of semantic values to simple content words that principally concerns us here, so it is worth considering how a possible worlds semantics can be and often is motivated by the claim that such a semantic theory enables us to provide adequate assignments of semantic values to simple content words.

Consider first an *extensional* semantic theory of the sort proposed in Irene Heim and Angelica Kratzer's (1997) introductory semantics textbook. On such a theory, we take the semantic values of names to be objects and the semantic values of predicates to be the the sets of objects to which those predicates apply.[2] It does not take long to see that extensional semantic values cannot possibly serve as adequate models of what speakers grasp in grasping the meanings of predicates. To see this, consider the following example. Suppose just three people smoke, Joe, Mary, and Sue. In which case, following two equations both specify the semantic value of "smokes":

$[\![\textbf{smokes}]\!] = \{x : x \text{ smokes}\}$
$[\![\textbf{smokes}]\!] = \{\text{Joe, Mary, Sue}\}$

Now, suppose it just so happens that Joe, Mary, and Sue are also the only individuals who ski. In which case, the following two equations both specifiy the semantic value of "skis"

[2] or the characteristic functions of such sets.

$[\![\textbf{skis}]\!] = \{x : x \text{ skis}\}$
$[\![\textbf{skis}]\!] = \{\text{Joe, Mary, Sue}\}$

As you can see, in this hypothetical scenario, the semantic value of "smokes" is identical to the semantic value of "skis." It is just the set of these three individuals. Clearly, however, in such a scenario, "smokes" would not mean the same thing as "skis;" "smokes" would still mean *smokes* and "skis" would still mean *skis*. It seems that our semantic theory ought not have the consequence that, if it just so happens that the same people who smoke are the ones who ski, the words "smokes" and "skis" would mean the same thing. Heim and Kratzer, of course, don't want their theory to have this consequence. In response to this problem, they say that, even though the expressions on the right of each of the two equations that are candidate specifications of the semantic values of "smokes" and "skis" define the same set, only the first type of expression is the sort that should go into our semantic theory. Only if we state the set that is the extension of the predicate with the use of a *condition*, specifying the set that is the semantic value of "smokes" or "skis" as the set of things that *smoke* or the set of things that *ski*, do we state that expression's semantic value in a way that "shows" its meaning. If we specify the set that is the extension of a predicate by merely listing its members, we do not show its meaning. The upshot of this response is to deny that the meaning of a predicate is to be identified with its semantic value. It is not sufficient, in the context of a semantic theory, to specify the semantic value of a predicate; one must specify it *in a particular way*, a way that "shows" its meaning.

In his discussion of the theoretical role that semantic values are to play, Yalcin takes issue with Heim and Kratzer's response to this problem and proposes an alternative:

> The conclusion to draw from the problem they raise is not that meaning must reside somewhere beyond semantic value; it is that the semantic values initially postulated are not fruitful, because too coarsegrained.

> We need richer semantic values to capture the sorts of distinctions that need distinguishing [. . .] A better response to the problem would simply be to introduce intensional resources (possible worlds or situations) at the start, as a beginning at fixing the problem, at delivering a semantic theory that can make at least a minimal range of the distinctions between semantic values that need to be distinguished.

Insofar as semantic values are the entities in theory that are meant to model the meanings of which speakers have knowledge, a theory that assigns the same semantic value to two predicates that differ in meaning but happen to have the same extension is inadequate. This inadequacy can be rectified, Yalcin suggests, by appealing to possible worlds from the outset in assigning semantic values to predicates. This alternative option is undertaken in several other introductory semantics textbooks, for instance Kate Kearn's textbook. There, she originally introduces the notion of possible worlds by considering the inadequacy of taking the extensions of predicates—the sets of objects to which they apply—to be their semantic values. She considers a theory that takes the semantic value of "brown" to be the set of brown things in the world, and she writes,

> Is that all there is to it? Suppose the world was exactly the way it is except for one detail—a certain brown pottery bowl on a windowsill in Ladakh is blue instead of brown. If the world was like that instead of how it is, then the set of brown things would be different, but surely the word *brown* wouldn't have a different meaning. This seems to make the word meaning depend on accidents of fate.
>
> We want to take into account the way the word *brown* would relate to the world even if things were a bit different from the way they actually are. We want to take into account not only the objects a predicate happens to apply to in fact, but also all the hypothetical objects that it would apply to, meaning what it does mean, if things were different. [. . .] We need to consider hypothetical versions of the whole of reality to state what individual predicates would apply to in virtue of their meaning. Words connect not only with the real world, but also with other possible worlds.

So, rather than taking the meaning of a predicate to be its *extension*, we can take the meaning of a predicate to be an *intension*, a function from possible worlds to extensions. The meaning of "brown" will thus be a function that maps each possible world to the set of things that are brown in that world. So, the meaning of "brown," in t he possible world Kearns describes, in which the pottery bowl of which she speaks is blue, is not different than it is in the actual world in which it is brown. In both possible worlds, "brown" still semantically expresses the function that maps each possible world to the set of brown things in that world. Though the set of brown things varies between the actual world and the possible world that Kearns describes, the function expressed by "brown" is invariant across these two possible worlds. We thus avoid the problem with the extensional theory of Heim and Kratzer, without having to say that meaning resides somewhere beyond semantic value.

The basic thought underlying this way of thinking about the meaning of a predicate is that to grasp the meaning of a predicate is to grasp a rule for sorting things into the things to which the predicate applies and the things to which it does not. More determinately, to grasp the meaning of a predicate is to grasp a rule for that enables one to take any possible way for things to be, any possible world, and sort things, as they are in that world, into two sets: the set of things to which the predicate applies and the set of things to which it does not. Accordingly, the meaning of a predicate can be modeled as a function that maps each possible world to the set of things that satisfy the predicate in that world. Thus, for instance, the meaning of "brown" can be modeled as a function that maps each world to the set of things that are brown in that world. Such functions are, at least in the context of such an extra-worldly framework, thought of as the properties that are expressed by those predicates. For instance, Paul Portner says, "a property can be thought of as an association between worlds and sets—it provides, for each world, a set of things." So, the property of being brown is thought of as being the

function that is the semantic value of "brown." What speakers grasp in grasping that "brown" and "pink" are incompatible is that, for any possible world, the set of things that are brown and the set of things that are pink are disjoint. No matter how things are, if something is pink, then it isn't brown, and vice versa.

Extra-worldly semantic theorists are often not particularly explicit as to whether properties, those entities that are semantically expressed by 1-place predicates, *really are* to functions from possible worlds to extensions, or whether properties are simply *adequately modeled by* functions from possible worlds to extensions, and likewise for the analogous question with respect to propositions and whether they are to be identified with sets of possible worlds (or functions from worlds to truth values). When theorists are explicit, their answers tend to vary. Andy Egan (2004), for instance, argues for *identifying* properties with functions from worlds to extensions. Other theorists, are explicit that they do not wish to make this identity claim. Commenting on the analgous question, Cheirchia and McConnel-Ginnet write

> We are not claiming that sets of worlds are what propositions really are. We claim only that sets of worlds have got the right structure to do what we think propositions do: mediate between sentences and their truth conditions. A crucial component of our semantic competence, what we understand when we understand the content of a sentence, is our capacity to match sentences and situations. Functions from worlds to truth values can be regarded as an abstract way of characterizing such a capacity.

When I get to my main argument in (§2.6), I will argue that, in order for our knowledge of the meanings of predicates to be *adequately modeled* by functions from worlds to extensions, what we actually know in knowing the meaning of a predicate must *really be* something quite close to such a function. It's important to be clear from the outset, however, that this is a claim that I'm making in the course of arguing against the *weaker* claim that functions from worlds to sets of objects cannot really be *adequate models* of what we grasp in grasping the meaning

of a predicates. To put this all in perspective, on the semantic theory I will eventually defend, the semantic value of a sentence will be a function that maps each discursive context in which someone is *entitled* to employ that sentence to the discursive context that would result upon their employing it. I do not claim that the meanings of sentences *really are* such functions, but I do claim that the meanings of sentences can be *adequately modeled* by such functions. That is what I am claiming is not so of the semantic values provided by the extra-worldly semantic framework.

2.3 A Simple Extra-Worldly Semantics

To investigate the structure an extra-worldly semantic framework more carefully and systematically, let us turn again to our toy language and lay out a simple extra-worldly semantic theory for it. We start with a model $\langle W, U, V \rangle$ consisting in a set of worlds W, a set of objects U, and a valuation function V. Let us consider first the first two elements of the model. In an extra-worldly semantic framework, we start with the notion of a completely determinate way for the world to be: a "possible world." The way the world actually is, of course, is one completely determinate way for the world to be. But there are different ways that the world could have been that are other than the way that it actually is. So, there are, we might say, other "possible worlds," worlds that are not actual, but merely possible. Applying this idea to the world in which our toy language is spoken and assuming, for the purpose of simplicity, that there can only be the three objects contained in it, there are twenty-seven completely determinate ways for the world to be—twenty-seven possible worlds. There is, for instance, the world in which a is gray, b is white, and c is black, there is the world in which a is gray, b is gray, and c is white, and so on. W is the set of these twenty-seven possible worlds, and U is just the set of the three objects contained in each of these

possible worlds, a, b, and c.

Now, consider the valuation function V. This function that assigns to a name a function that maps each possible world to a particular object (the one that is actually named by that name), assigns to a 1-place predicate a function that maps each possible world to a set of objects (the ones that satisfy the predicate in that world), and assigns to a 2-place predicate a function that maps each possible world to a set of pairs of objects (the pairs that satisfy the predicate in that world). So, our valuation function will assign to the the name "a" a function that maps each possible world to a, it will assign to the predicate "is grey" a function that maps each possible world to the set of things that are grey in that world, it will assign to the predicate "is darker than" a function that maps each possible word to the set of pairs of objects such that the first is darker than the second in that world, and so on. It thus assigns meanings to the basic content words of the language for which we are constructing a semantic theory; it tells us what objects are named by the names of our toy language, which properties are expressed by the 1-place predicates of our toy language, and which relations are expressed by 2-place predicates of our toy language. This constitutes the ground level of the semantic theory.

Having specified a model of this sort, our aim in constructing a semantic theory is to devise a way of assigning semantic values to all of the complex expressions of our language on the basis of the valuations that our model assigns to simple ones. So, for name n, and for a 1- or 2-place predicate P, the value is just what our model gives us:

1. $[\![n]\!] = V(n)$
2. $[\![P]\!] = V(P)$

Once we've specified these values, we can assign a value to any sentence consisting in a name followed by a 1-place predicate, and a name followed by a 2-place

predicate and then another name as follows:[3]

1. $[\![nP]\!] = \{w : [\![n]\!](w) \in [\![P]\!](w)\}$
2. $[\![n_1Pn_2]\!] = \{w : ([\![n_1]\!](w), [\![n_2]\!](w)) \in [\![P]\!](w)\}$

So, some world w is an element of $[\![nP]\!]$ just in case the object to which $[\![n]\!]$ maps w is an element of the set of objects to which $[\![P]\!]$ maps w. And some world w is an element of $[\![n_1Pn_2]\!]$ just in case the pair of objects consisting in the object to which $[\![n_1]\!]$ maps w and then the object to which $[\![n_2]\!]$ maps w is an element of the set of pairs of objects to which $[\![P]\!]$ maps w. The result of these composition rules is that atomic sentences of our toy language are assigned some subset of these possible worlds as semantic values. Consider, for instance, the set of worlds that will be assigned to "a is gray." The semantic value of a is a function that maps each possible world to a, and the semantic value of "is grey" is a function that maps each possible world to the set of gray things in that world. The semantic value of "a is grey," then, will be the set of worlds such that a is an element of the set of grey things in those world. That is it, will be the set of worlds in which a is gray. Likewise, the semantic value of the sentence "a is darker than b" will the set of worlds in which a is darker than b, the semantic value of the sentence "a is the same color as a" is the set of all twenty-seven possible worlds, since every world is such that each object in it is the same color as itself, the semantic value of the sentence "a is lighter than a" is the set of no possible worlds, and so on.

Once we've assigned values for all the atomic sentences in this way, we can assign values to logically complex sentences with the use of set-theoretic operations of complementation, intersection, and union as follows:

2. $[\![(\textbf{It is not the case that } \phi)]\!] = ([\![\phi]\!])'$
3. $[\![(\phi \textbf{ and } \psi)]\!] = [\![\phi]\!] \cap [\![\psi]\!]$

[3]This notation—specifically, using functional notation of the form $f(x)$ with semantic values as the functions and worlds as arguments—is non-standard. I have put things this way for simplicity.

4. $[\![(\phi \textbf{ or } \psi)]\!] = [\![\phi]\!] \cup [\![\psi]\!]$

Assigning values to complex sentences in this way, we have a semantics with which we can assign semantic values to all of the sentences of our toy language, all infinity of them. So, for instance, the set of worlds assigned to "It's not the case that *c* is gray," will be the set of all the worlds that are not elements of the set of worlds in which *c* is gray. The set of worlds assigned to "*a* is lighter than *b* or *b* is white" will be the set of worlds that are either an element of the set of worlds in which *a* is lighter than *b* or an element the set of worlds in which *b* is white. The set of worlds assigned to "(It's not the case that *c* is gray) and (*a* is lighter than *b* or *b* is white)" will be the intersection of the first set and the second set. And so on.

We now have a function $[\![\cdot]\!]$, defined for all of the sentences of our toy language, such that, for any sentence of our toy language ϕ, $[\![\phi]\!]$ is a formal model of the meaning of ϕ. We have a simple semantic theory for our toy language. Now that we've constructed a simple semantic theory, let's see what theoretical work it can do. Consider again the following set of facts:

F1. The sentence "*a* is darker than *b*" is synonymous with the sentence "*b* is lighter than *a*."

F2. The sentences "*a* is black" and "*b* is white" jointly entail the sentence "*a* is darker than *b*."

F3. The sentence "*a* is black" is incompatible with the sentence "*a* is white."

As we've said, these are the sort of facts for which we want our semantic theory to account. Our guiding idea in constructing a semantic theory that can account for these facts is that speakers of a language behave in certain ways because they know that certain sentences of their language are synonomous with one another, entail one another, or are incompatible with one another, and they have this knowledge because they know what these sentences of mean. If, by assigning meanings

to these sentences, our semantic theory enables us to account for facts like (F1)-(F3), then, by taking speakers to have knowledge of meanings we assign to these sentences, we can explain their knowledge of these facts, thereby explaining their behavior as a manifestation of this semantic knowledge. The simple possible worlds semantics we've just sketched promises to enable us to do this. Let's see how.

On the simple semantic theory just sketched, we can give the following definitions. First, two sentences, ϕ and ψ, are synonomous just in case the set of worlds that is the value of ϕ is identical to the set of worlds that is the value of ϕ. That is, ϕ is synonomous with ψ just in case $[\![\phi]\!] = [\![\psi]\!]$. So, "$a$ is darker than b" is synonomous with the sentence "b is lighter than a" just in case every world that is an element of the set of worlds in which "a is darker than b" is true is also an element of the set of worlds in which "b is lighter than a" is true, and vice versa. Since that is indeed so, (F1) obtains. Second, two sentences, ϕ and ψ, jointly entail another sentence, ξ, just in case the intersection of the sets of worlds that is the value of ϕ and the set of worlds that is the value of ψ is a subset of the set of worlds that is the value of ξ. That is, two sentences ϕ and ψ jointly entail another sentence, ξ just in case $([\![\phi]\!] \cap [\![\psi]\!]) \subseteq [\![\xi]\!]$. So, "$a$ is black" and "b is white" jointly entail "a is darker than b" just in case any world that is an element of both the set of worlds in which "a is black" is true and the set of worlds in which "b is white" is true is an element of the set of worlds in which "a is darker than b" is true. Since this is so, (F2) obtains. Finally, two sentences ϕ and ψ are incompatible just in case the sets of worlds that are their values are disjoint. That is, ϕ is incompatible with ψ just in case $[\![\phi]\!] \cap [\![\psi]\!] = \varnothing$. So, "$a$ is white" is incompatible with "a is black" just in case there is no world that is an element of both the set of worlds in which "a is white" is true and the set of worlds in which "a is black" is true. Since there is no such world, (F3) obtains. With these definitions, it seems that the simple possible worlds semantics just sketched enables us to account for (F1), (F2), and (F3).

Given that we can account for (F1)-(F3), it seems that, by modeling speakers' knowledge of the meaning of the sentences "*a* is white," "*a*" is black," "*a* is darker than *b*," and "*b* is lighter than *a*" as knowledge of the semantic values that our semantic theory assigns to them, we can explain speakers' knowledge of these facts, thereby explaining the behavior they exhibit in virtue of having this knowledge. Consider, for instance, the fact that competent speakers of our toy language behave in way that manifests their knowledge of the fact that sentences "*a* is white" and "*a* is black" are incompatible. Recall, they never utter both sentences at the same time, they correct incompetent speakers that do, and so on. The explanation of this behavior, on this model, is that they know that the sets of worlds that are the values of these two sentences are disjoint. Uttering "*a* is white" would function to inform other speakers that the actual world is among the set of worlds in which *a* is white. Uttering "*a* is black" would function to inform other speakers that the actual world is among the set of worlds in which *a* is black. The knowledge of the incompatibility of these two sentences that competent speakers have consists in their knowledge that the sets of worlds that are the values of these two sentences are disjoint. Having this knowledge, they know that uttering both sentences would function rule out every possible world. Knowing this, they know to never utter both sentences at the same time, to correct incompetent speakers that do, and so on. In this way, it seems that our simple extra-worldly semantics enables us to explain the behavior we set out to explain. Things, however, are not how they seem. To see why this is so, let us turn to the core notion of a possible worlds semantics: the possible world.

2.4 The Issue of Defining Possible Worlds

In a possible worlds semantics, a possible world w is often officially defined as a function that maps each sentence in the set of atomic sentences $\mathcal{A}$ to one of

two values, *true* or *false* (Dever 2012, 51; Willer 2013, 9; Goldstein 2018, 4). There are other formally interchangeable definitions, but I'll call this one the "standard definition."[4] Officially, the standard definition is the following:

A possible world w is any function $f : \mathcal{A} \rightarrow \{true, false\}$.

The intuition behind this definition is clear enough. A possible world is something that determines, for each atomic sentence of the language, whether that sentence is true or false. Accordingly, a possible world w can be defined as a function that maps each atomic sentence to a value, *true* or *false*. Having defined possible worlds as these functions, we can officially say what it is for an atomic sentence to be true in a world as follows:

For any atomic sentence p, p is true in w iff $w(p) = true$

This enables us to officially assign truth values to atomic sentences relative to possible worlds at the base level of our semantic theory, and then we can go from there.

Though the standard definition is widely treated as good enough for the purposes of laying down the groundwork for a possible worlds semantics, it does not take long to see what is wrong with it. Not only does it give us possible worlds, but it gives us "worlds" that are not possible as well. For instance, "a is white" and "a is black" are both atomic sentences, and so there is a function $f : \mathcal{A} \rightarrow \{true, false\}$ that maps "a is white" to *true* and maps "a is black" to *true*. On the standard definition, this gives us a "possible world," one in which "a is white" is true and "a is black" is true. But clearly there is no possible world in which "a is white" is true and "a is black" is true. If a is white, then it can't be the case that a is black. So, there is no possible world in which "a is white" is true

[4]We could equally define a possible world as a subset of the set of atomic sentences $\mathcal{A}$, and then the standard definition would be the characteristic function of that set (Veltmann 1996, 228)

and "*a* is black" is true. But standard definition says there is. That's a problem. Though this problem is completely obvious, it turns out to be critical.

A first response is to offer a revised definition. One might, for instance, start by specifying which sets of atomic sentences are incompatible and then say that a possible world is a function that maps each atomic sentence the language to a value *true* or *false* in such a way that it does not map all the members of any such set to the value *true*. This excludes a function that maps both "*a* is white" and "*a* is black" to *true* from being a possible world, since the set consisting of "*a* is white" and "*a* is black" is a set of incompatible sentences. The problem with saying this, however, is that our simple semantic theory was supposed to enable us to *account for* the fact that the sentence "*a* is black" is incompatible with the sentence "*a* is white." Defining possible worlds in such a way that they depend on this fact precludes us from being able to do this. As we've said, on the simple semantics we've given, two sentences ϕ and ψ are incompatible just in case $[\![\phi]\!] \cap [\![\psi]\!] = \varnothing$. So, "*a* is white" is incompatible with "*a* is black" just in case there is no world that is an element of both the set of worlds in which "*a* is white" is true and the set of worlds in which "*a* is black" is true. Is there any such world? Well, on the standard definition there is; on the revised definition, there is not. However, the reason *why* there is no such world on the revised definition is that "*a* is white" is incompatible with "*a* is black," so the revised definition does not count any function that maps both sentences to the value *true* as a world. Since the fact that "*a* is white" is incompatible with "*a* is black" *explains why* there is no world that is an element of both the set of worlds in which "*a* is white" is true and the set of worlds in which "*a* is black" is true, saying that the two sets are disjoint cannot amount to giving an account of *what it is* for "*a* is white" to be incompatible with "*a* is black." Schematically, if the fact that *A explains* the fact that *B*, the fact that *A* cannot *just consist in* the fact that *B*. This is the first instance of a principle to which we will return several times throughout this dissertation. Here, the upshot should

be quite clear: If we adopt the revised definition, the "account" of incompatibility given by possible worlds semantics cannot be an account at all.

A more sophisticated response to our problem, owed to Rudolph Carnap (1956) and advocated in contemporary semantics by Barbara Partee (2005), is to say that, in order to properly define the space of possible worlds, we must lay down certain "meaning postulates" which function to constrain the model on which we base our semantic theory so that it includes only genuinely possible worlds. We said that our valuation function, applied to a 1-place predicate, gives us a function that maps each possible world to the set of objects that satisfy that predicate in that world. What we need to do is restrict the set of possible worlds that we let into our model by specifying, which predicates can't be jointly satisfied by a single object in a world, which predicates which, if they are satisfied by some objects, require other predicates to be satisfied, and so on. Here, according Carnap and Partee, is where "meaning postulates" are meant to come in. The idea is that if we want to constrain which "worlds" get included in the model on which we base our possible worlds semantics, we can do this by laying down something like the following:

$$\forall x(\textbf{white}(x) \rightarrow \neg \textbf{black}(x))$$

Here, **white** is the symbol that we're using in our semantic theory to symbolize the predicate "is white" of our toy language, and **black** is the symbol that we're using for the predicate "is black."[5] This postulate says that, for any object x, if x satisfies the predicate "is white," then it is not the case that x satisfies the predicate "is black." Laying down this postulate enables us to put a constraint on which "worlds" get counted as worlds in our model—it enables us to rule out any "world" w in which there is some object x, such that x is an element of $V(\textbf{white})(w)$ and an element of $V(\textbf{black})(w)$. We are thus able to formally capture the fact that

[5]Carnap (1952).

the predicates "is white" and "is black" are incompatible in our semantic theory by making it such that the model on which it is based contains no world w in which there is an object x that satisfies both predicates.

It is now crystal clear, however, that our possible worlds semantics, based on a model that is determined by meaning postulates of the above sort, does not and cannot give us an account of the fact that the predicate "is white" and the predicate "is black" are incompatible or our knowledge thereof. Our semantic theory only captures this fact because we laid down the meaning postulate that we did, and we laid down this meaning postulate only because we know that the predicate "is white" and the predicate "is black" are incompatible. So, since the our knowledge of that fact that "is white" is incompatible with the predicate "is black" *explains why* our semantic theory contains no world in which there is some object x that satisfies both predicates, we cannot *account for* this fact or our knowledge thereof with the use of this semantic theory. This is essentially the same issue as with the revised definition; it just arises here at the level of predicates rather than the level of sentences. In order to define possible worlds, we must appeal to our knowledge of the very facts that our possible worlds semantics was meant to explain. Of course, if our aim is just to elucidate our semantic knowledge rather than to account for it, then there is no problem here. As discussed earlier, however, contemporary semantics, at least as it is often advertised, has more than merely elucidatory ambitions.

It's worth noting that Carnap himself is quite clear about the fact that he is not aiming to account for speakers' knowledge of meaning. Commenting on what grounds the theorist's writing down certain meaning postulates, he writes,

> How does [the theorist] know that these properties are incompatible and that therefore he has to lay down postulate P_1? This is not a matter of knowledge but of decision. His knowledge or belief that the English words 'bachelor' and 'married' [or 'white' and 'black' in our case] are always or usually understood in such a way that they

> are incompatible may influence his decision if he has the intention to reflect in his system some of the meaning relations of English words.

Here, Carnap says that if we want our system to *reflect* the fact that certain words "are always or usually understood in such a way that they are incompatible," we'll lay down certain meaning postulates rather than others. He is under no illusion that a semantic theory of the sort he is proposing will be able to *account for* the understanding of the incompatibility of certain English words that English speakers have; it will simply reflect this understanding. In other words, Carnap's ambitions here are self-consciously *elucidatory* rather than *explanatory*. In *Meaning and Necessity*, he's clear that his main aim in providing the semantic analyses that he does is the *clarification* of philosophical concepts, aiming to replace the vague concept of, say, a sentence's being necessarily truth, with the precise concept of a sentence's holding in every state description (1947, 7-13). Carnap's aims, however, are not those of contemporary linguistic theorists, who do take themselves to be engaging in a genuinely explanatory enterprise.

Insofar as our aims are more than merely elucidatory, possible worlds cannot be defined in terms of sentences or predicates of the language for which one is constructing a semantic theory, for doing so requires one to the very semantic knowledge that is supposed to be accounted for by the theory. I take it that, by and large, those who employ talk of possible worlds with genuinely explanatory ambitions will not resist this conclusion. By such a theorist's lights, a possible world is simply a completely determinate way for the world to be. The set of possible worlds that there is does not depend on the semantic relations that obtain between expressions of a language, but, rather, simply on the set of possible worlds that there really are or on the set of ways that the world can possibly be. Knowledge of possible worlds thus does not require semantic knowledge of the sort that a possible worlds semantics seeks to explain. Rather, it is simply knowledge of the set of possible worlds that there really are or of the set of ways

that the world can possibly be. So, the knowledge that underlies the knowledge of the fact that the sentence "*a* is grey" is incompatible with sentence "*a* is black" is either the knowledge that there is no world in which *a* is gray and *b* is black or that the world cannot be such that *a* is gray and *a* is black. That's not a fact about meaning but a fact about the world or, perhaps, the worlds.

Now, at this point, once one takes knowledge of possible worlds to be knowledge of non-linguistic worldly entities rather than knowledge of linguistic entities and their semantic relations, there are two ways to go: one can take possible worlds to be *composed* out of other worldly entities, such as propositions, states of affairs, or properties, as Adams (1974) and Plantinga (1976) classically do and Soames and King more recently do, or one can take possible worlds to be *primitive* worldly entities. For reasons that will become clear shortly, if they aren't clear already, I will put off discussion of the former proposal for the next chapter, which concerns intra-worldly knowledge, and consider here the primitivist proposal, which takes knowledge of possible worlds to be a basic sort of worldly knowledge, not derivative on intra-worldly knowledge. An extra-worldly semantics proper, the sort of semantic theory proposed by Lewis and Stalnaker, takes our semantic knowledge to be based on worldly knowledge that is properly extra-worldly.

2.5 The Primitivist Proposal

The definitions of possible worlds that we've just considered aim to define possible worlds in terms of expressions of the language for which we're giving a possible worlds semantics. David Lewis, one of the philosophical pioneers of extra-worldly semantics, rejects this sort of approach, and he does so largely because of the issue with which we're concerned. He writes,

> [I]t would do us nothing to identify possible worlds with sets of sentences (or the like), since we would need the notion of possibility oth-

> erwise understood to specify correctly which sets of sentences were to be identified with worlds, (Lewis 1973, 86).

Lewis recognizes here, that, if we identified possible worlds with formal constructions from sentences, we could not then use a semantics based on possible worlds in order to give an account of the notion of possibility. In order to say which sets of sentences are to be identified with worlds, we'd need to say which sets of sentences are compossible, and to do that would be to appeal to the very modal notions that we're trying to account for with the use of possible worlds. Now, Lewis's principle concern is with giving an semantics for modal sentences, but the same issue applies in our attempt to account for the modally robust semantic relations that obtain between ordinary non-modal sentences. Though our example has led us to focus on the notion of *in*compossibility here—the incompatibility of the sentences "*a* is white" and "*a* is black"—the issue is just the same. It is the issue that precludes us from being able to use any of the accounts of possible worlds considered thus far if we're going to attempt to employ a possible worlds semantics to give an account of what it is for two sentences to be compossible or incompossible—compatible or incompatible. That's our issue. Lewis proposes a novel solution to it. Rather than thinking of possible worlds as sets of sentences or functions from sentences to truth values, Lewis opts to think of them as particular objects just like the actual world (Lewis, 1973, 1986).

Lewis's approach is a "primitivist" one: we do not try to say what a possible world is in terms of entities that are not possible worlds. By Lewis's lights, we know what sort of thing the actual world is, and that's a possible world, so it's sufficient to say that other possible worlds are other entities just like the actual world. To say that this world is the "actual" one, on his view, is not to claim that there is a special property of existence or reality that only this world has. For Lewis, "actual" doesn't mean existent or real. Rather, he thinks of it as an indexical like "here" or "now." Just like "now," when uttered by some speaker at

some time, just picks out the time at which one happens to be speaking, "actual," when uttered by some speaker in some world, just picks out the world in which one happens to be speaking.[6] Just as thinking of "now" as an indexical opens the door for the view that times other than the one we happen to be in are equally real, thinking of "actual" as an indexical opens up the door for genuine realism about possible worlds, the view that possible worlds other than the one that we happen to be in are equally real. Having opened this door, Lewis, in what can only be described as an act of intellectual bravery, walks through it.

So, Lewis's way of being a primitivst about possible worlds is to be a genuine realist about them. Most theorists, however, have not wanted to walk through this door with Lewis. Perhaps the most prominent such theorist is Robert Stalnaker, another philosophical pioneer of extra-worldly semantics. Stalnaker (1986), endorsing what he calls "modest realism," maintains with Lewis that possible worlds aren't to be defined in terms of things other than possible worlds, but he does not go all the way to the genuine realism of Lewis. The view starts with the thought that there are two ways in which one might take there to be "many ways things could have been besides the way they actually are," (Lewis, 1973). On the one hand, one might take this world to be what we're speaking of when we speak of "the way that things actually are," and take there to be other worlds, other entities just like this one, that we are speak of when we speak of "other ways that things could have been." On the other hand, one might think that what we're speaking of when we talk of "the way things actually are" isn't the actual world *itself*, but *the way the actual world is*: the property that the actual world instantiates in being just the way that it actually is. This, Stalker thinks, is an important

[6] "This world is actual," is kind of like saying "I am me." Accordingly, saying "This world is different than all the other possible worlds—only this one is *actual*" is like saying "But I'm different than all the other people—only I am *me*." To say this and think one is making a substantive claim about a special property that one has would be to misunderstand the logic of indexical language. If this is the correct way of thinking about what "actual" means, then we should not take the claim "This world is the only one that is actual" to entail "This world is the only one that is real." To think that actuality and reality had to go together would be analogous to solipsism.

distinction. Making it, we are able to maintain that there ways that things could have been, without thinking that they are the same sort of thing that the actual world is. Possible worlds aren't concrete objects, but abstract objects: properties, the sort of thing that objects instantiate, rather than objects themselves (objects, that is, which are not properties). While the actual world itself is not a property, the way the actual world is *is* a property, the property that the actual world instantiates in being just the way that it is. If the actual world were some other way, then, being this other way, it would instantiate some other property, some other way for the world to be. On the Stalnaker's view, there actually are all of these other properties; they are "possible worlds" (or, perhaps less misleadingly called, "possible world-states") that figure into our semantic theory.

Before I go on to argue against the application of these primitivist conceptions of possible worlds in our semantic theory, I want to briefly consider one property that Stalnaker thinks possible worlds might reasonably be taken to have that makes them better candidates than Lewis's genuine other worlds. Stalnaker takes it that his worlds can reasonably be taken to be such that their existence depends on the activities of language speakers. On the basis of saying that that possible worlds are "abstract objects whose existence is inferred or abstracted from the activities of rational agents," Stalnaker says "It is thus not implausible to suppose that their existence is in some sense dependent on, and that their natures must be explained in terms of, those activities, (51-52)" This suggestion of Stalnaker's is cited approvingly by many theorists of meaning who wish to employ the framework of possible worlds in order to give an account of the meanings of the expressions of a natural language without having to commit themselves a potentially scientifically questionable sort of realism about these entities (Cheirchia and McConnell-Ginnett 1990, 207-208; Partee 1988, 102). I take, however, that this suggestion of Stalnaker's, in conjunction with the explanatory aim of semantics, is incoherent. One place where this incoherence arises is in

Gennaro Chierchia and Sally McConnell-Ginnet's (1990) introductory textbook in formal semantics. Consider first how they articulate their project in linguistics from the outset:

> The linguistic knowledge we seek to model, speakers' competence, must be distinguished from their observable linguistic behavior. Both the linguist and the physicist posit abstract theoretical entities that help explain the observed phenomena and predict further observations under specified conditions, (2).

Possible worlds are precisely the sort of "abstract theoretical entities that help explain the observed phenomena and predict further observations" that Chierchia and McConnell Ginnet have in mind here. When they end up introducing these entities, they reaffirmingy say, with explicit reference to the Stalnaker passage quoted above,

> The human activities on which the existence of possible worlds depends (through which they are stipulated) include using language and interpreting expressions. Semantics, as we understand it, seeks to develop a model (or part of a model) of such activities, (207).

On the face of it, what they say here might seem to be in line with what they say in the quote above it about positing theoretical entities to model semantic competence and thereby explain linguistic activities. However, though the above two quotes may seem to be compatible, there is a tension here. One the one hand, they claim that possible worlds depend for their existence on the linguistic activities that are a manifestation of semantic competence, "using language and interpreting expressions." On the other hand, they claim that these linguistic activities are to be explained by a semantic theory that features possible worlds. One cannot coherently maintain both of these claims at once.

Consider again the example from our toy language. The set of activities that are a manifestation of our speakers' semantic competence includes their acting in such a way that shows that they understand the sentences "*a* is white" and "*a* is

black" to be incompatible. They never utter both sentences at the same time, they correct incompetent speakers that do, and so on. These activities, we theorize, are manifestations of their knowledge of the meanings of the sentences "*a* is white" and "*a* is black." Now, suppose we posit a class of theoretical entities that we call "possible worlds" in order to say what it is in which their knowledge consists. We say that their knowledge consists in their grasp of a particular fact about two sets of possible worlds, the ones that we theorize to be, due to their linguistic conventions, the correspondents of "*a* is white" and "*a* is black." It is in virtue of knowing that these two sets are disjoint and knowing the conventions of their langauge that they know that one cannot correctly utter both sentences at the same time, and this knowledge explains why they act as we do, never uttering both sentences at the same time, correcting others that do, and so on. In this way, we explain their activity by taking them to bear a cognitive relation to two sets of entities of the sort that we've posited—we explain their activity by saying that they know a certain fact about two sets of possible worlds, the set of worlds in which *a* is white and the set of worlds in which *a* is black. If this is the form of our explanation of their linguistic activities, the existence of these possible worlds can't depend on the activities that they are posited to help explain. If they were so dependent, they wouldn't be able to figure into the explanation of these activities in the way that they do. So, if we take a possible worlds semantics to be able to give an account of the knowledge that they have in knowing the meanings of the sentences of their language, and we take this knowledge to explain their linguistic activities, we can't take the existence of possible worlds to depend on these activities.

Once we drop Stalnaker's suggestion that possible worlds depend on the linguistic activities of speakers, I take it that it does not matter much whether we follow Lewis in taking possible worlds to be concrete objects like the actual world or follow Stalnaker in taking them to be abstract properties like the property that

the actual world instantiates in being just the way it is. Whatever we say—whether we say that other possible world are "other things" like the actual world or "other ways" like the way the actual world is—the basic picture is roughly the same: there is a space of primitive possibilities whose existence does not depend on us, to which we have cognitive access, and this cognitive access is a precondition for the possession of contentful mental states and semantic knowledge. This picture, whether Lewisian or Stalnakerian, is what I'll call the "primitivist picture."

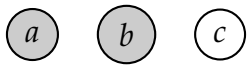
For the purposes of the present discussion, I will simply grant the cognitive access to worlds that we are supposed to have on the primitivist picture. Much ink has been spilled over the problem of our cognitive access to other possible worlds, be they concreta or abstracta. For my purposes here, I simply grant here that we have this access. The claim that I am about to make is that, even with this access being granted, trying to construct a semantic theory on the basis of such access commits one to a fatal instance of the Myth of the Given.

2.6 The Myth of the Extra-Worldly Given

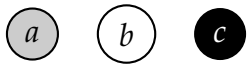
The problem I am about to raise for extra-worldly semantics can be raised with respect to both propositions and properties. I will raise it just for the extra-worldly conception of properties here, since, given the compositionality of meaning, this suffices to raise a problem for the extra-worldly conception of propositions, and, as a result, for the whole theory. Let me first restate the extra-worldly conception of properties, which either are or are modeled by the semantic values assined to (1-place) predicates.

On the extra-worldly conception of meaning, to grasp the meaning of a predicate is to grasp the property expressed by that predicate, where this property, modeled as a function from worlds to extensions, is thought of as something such that the grasp of it enables you to relate any possible world to a particular set of

things in that world. For instance, the property of being gray is something such that grasping it enables you to take any possible world and single out set of things that are gray in that world. So, if one grasps the property of being gray, then, given the following possible world:



one is able to single out the set $\{a, b\}$. Given the following possible world,



one is able to single out the set $\{a\}$, and so on for every possible world. Given that properties serve this cognitive role, they can be modeled as functions that map possible worlds to extensions. To have a grip on such a function would be to have a grip on exactly the sort of thing that would enable you to go from a possible world to the set of things that instantiate the property in that world. The theoretical role of properties, modeled by such functions, is that our grip on them explains this ability to, no matter how things are, sort things into the things that instantiate them and the things that do not. That's the thought.

The thought may all seem well and good, but I take it that there is a serious problem with it. I'll make this problem explicit by way of the following argument:[7]

1. One's grasp of a property can be adequately modeled by one's grasp of a function from worlds to extensions only if one's grasp of that property is identified with one's grasp of a rule/mapping that takes one from worlds to extensions, (premise).
2. One's grasp of and ability to ascribe a property explanatorily grounds one's grasp of a rule/mapping that takes one from worlds to extensions, (premise).
3. If some thing A explanatorily grounds some other thing B, A cannot be identified with B, (premise).

[7]Note, when I say "property" in the context of this argument, I mean to be speaking of *simple material* properties, like the property of being gray.

4. So, one's grasp of a property cannot be identified with one's grasp of a rule/mapping that takes one from worlds extensions, (2, 3).

5. So one's grasp of a property cannot be adequately modeled by one's grasp of a function from worlds to extensions (1,4).

This argument is clearly valid. To see whether it is sound, let's go through it premise by premise.

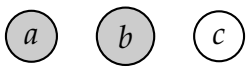
First, let us consider (1). Proponents of extra worldly semantics will often speak of grasping a meaning as grasping a "rule" taking one from worlds to extensions.[8] I write "rule/mapping" in (1) because the notion of a "rule" that is employed in this context can be nothing other than a (perhaps normatively expressed) mapping, something that takes one from a given world to an extension in that world.[9] A "rule," in this sense, is something that "tells one" (either literally or metaphorically, but presumably metaphorically): In case A, do x; in case B, do y; and so on. When Stalnaker, for instance, speaks of understanding a sentence, grasping a the proposition expressed by it, as knowing a "the rule for determining the truth value of what was said, given the facts," he cannot be speaking of anything other than knowing a mapping from the various possible situations to a truth value. If this were not what he was speaking of, there would be no reason to think, on the basis of this explication, that grasping a proposition could be thought of as the grasp of a function from worlds to truth values. Likewise, in the case with which we are presently concerned. In order to think that what one grasps in grasping a property is adequately modeled by a function from worlds to extensions, one must think that what one grasps, in grasping a property, is a rule for determining the set of things that instantiate that property, given the facts,

[8]

[9]Note here that there is a difference between the intuitive notion of a "mapping" that I am using to explicate the use of the term "rule" here and the technical notion of a *function*. Following standard practice in formal semantics, I use "function" in its mathematical sense: a function f is a set of ordered paris that such that, for any x, y, and z, if $(x, y) \in f$ and $(x, z) \in f$, then $y = z$. A rule or mapping taking one from possible worlds to sets of objects is *modeled by* a set of ordered pairs that has, as first elements, possible worlds, and, as second elements, sets of objects.

where a "rule" here, is simply a mapping. That is how we've been talking this whole time, and way of speaking reflects the way of thinking that underlies this way of trying to model the meanings of predicates.

Now let us consider (2), which is the crucial, though I believe obvious, premise. Note first that, once again, I am simply assuming that we have cognitive access to possible worlds. We are, somehow, capable of getting various possible worlds "into view," in whatever, presumably metaphorical, sense in which we are supposed to be able to have them "in view." The question is: *how is it* that we are able to take any possible world that we have "in view" and single out the set of gray things in that world? Grasp of a mapping, of course, *does* enable us to go from any possible world to the set of things that are gray in that world, for, given any possible world, such a mapping simply "tells us," which things are gray in that world. But to answer the question of how it is that we are able to take any possible world and single out the set of gray things in that world by saying "We grasp a mapping taking any possible world to the set of gray things in that world" is not to answer the question at all. It is, in response to the question of *how* we are able to take any possible world and single out the set of gray things in that world, to say *we just are*. But there *is* an answer to the question of how it is that we are able to take any possible world and single out the set of gray things in that world. Indeed, there is an *obvious* answer. The answer to the question of how it is that we are able to take any possible world and single out the set of gray things in that world is that we know what it is for something to be gray and are capable of recognizing things as being such. That is, we grasp the property of being gray and, given any possible world that we have "in view," we are capable of ascribing it to gray things in that world. For instance, given the world



we are able to single out the set $\{a, b\}$ because we know what it is for something

to be gray and we are capable of recognizing *a* and *b* as being such. That is, we grasp the property of being gray and we are capable of recognizing *a* and *b* as instantiating this property. This is obvious, but the possible world theorist is unable to say it. They are unable to say it because it appeals to our grasp of the property of being gray in order to explain our grasp on a mapping from worlds to extensions. If our grasp of the property of being gray is taken to *just be* this mapping, we cannot appeal to the the former grasp to explain the latter.

Substantiating this last claim brings us to premise (3), which will be familiar from our discussion of the revised definition of possible worlds. I take this schema to be a statement of what is fallacious in the Euthyphro fallacy, the relevance of which in semantics has been brought to attention by Jason Bridges (2006).[10] When asked what it is for someone to be pious, Euthyphro claims that for someone to be pious is for them to be an object of God's love. However, when asked why it is that God loves the people that he does, Euthyphro claims that it is because they are pious. This pair of claims is inconsistent. If someone's being pious *explains* their being an object of God's affection, their being pious cannot be *identified with* their being an object of God's affection. (3) is a generalization of this fact, schematizing the explanans and explanandum that are not to be identified as *A* and *B* respectively. It's at all not clear that (3) needs or is capable of being given a justification from a principles clearer than itself, but, if it does help, we can note that the truth of (3) follows from the fact that explanation is an asymmetric relation, whereas identity, if it is a relation at all, is clearly a symmetric one.

(4) follows from (2) and (3), and (5) follows from (1) and (4).[11] I conclude that an extra-worldly semantics is incapable of adequately modeling our knowledge

[10]For an interesting discussion of how this fallacy plagues certain causal/informational conceptions of semantics, see Bridges (2006).

[11]I take it that the logical structure here is quite transparent, and I'd be extremely surprised if any response to this argument involved rejecting any of the logical principles used in drawing this conclusion. For completeness, however, the first step is a universal instanatiation, and the second step is either a modus ponens or tollens, depending on whether the "only if" specification of a conditional is formalized as a contrapositive or not.

of the meanings of predicates.

It is clear what the issue is here. One's grasp of a propertie explains one's ability to grasp functions going from worlds to extensions. Indeed, one's grasp of properties explains one's ability to grasp possible worlds at all. The theorist can only define these entities, bringing into her universe possible worlds on which speakers are supposed to have a grip, insofar as she populates them with things whose properties she is capable of recognizing. She thus must draw on her capacity to recognize gray things and to ascribe the property of being gray to them in order to bring into view the entities in terms of which this capacity is supposed to be analyzed. She must suppose that the availability of possible worlds for cognition does not draw on the capacities that are in fact required for this cognition. Accordingly, if her theory is not to be incoherent, she must suppose that the structure of her semantic universe simply imposes itself on the mind, that the worlds that populate it are organized set-theoretically, such that, given a world, one can immediately pick out the sets of things in that world, as organized by the predicates of the language for which she is constructing a semantic theory. But, in that case, the knowledge that is supposed to underlie speakers' knowledge of meaning becomes utterly unnintellegible.

Ultimately, I take it that the extra-worldly semanticist, insofar as they maintain their commitment to extra-worldly semantics as an explanatory enterprise, will be forced to accept one of two horns of a dilemma concerning our knowledge of extra-worldly semantic values: either our knowledge of them is *unintelligible* or it is *incoherent*. This is the basic dilemma faced by anyone who is committed to an instance of the Myth of the Given. Extra-worldly semantics, as I've spelled it out here, is clearly an instance of the Myth. The structure of the world—understood here in terms of the set-theoretic organization of the possible worlds that populate the semantic universe—is taken to impose itself on the mind from without. The problem with the impositionist picture here is that, given the structure of the

theory, our grasp on this structure must not draw on the capacities that are, in fact, required to grasp it, specifically, the capacity to recognize things as instantiating properties, ascribing those properties to them. If one refuses to recognize the capacities that this grasp in fact requires, then the resultant theory is one in which our grasp on this structure is *unintelligible*; we are simply taken to have this grasp without there being any explanation of how we do. If one does recognize the capacities that this grasp in fact requires, then the resultant theory is *incoherent*; this grasp is taken to not require the capacities that it is acknowledged that it does, in fact, require.

2.7 The Problem Percolates Up

This basic problem critically infects the whole semantic theory. Consider first, (F3), the fact "*a* is grey" is incompatible with "*a* is white." The explanation of this fact, on this semantic theory, is that the predicates "is grey" and "is white" are incompatible, and this theory purports to give us a way of modeling this incompatibility. If properties are modeled as functions from worlds to extensions, then we can also explain our grip on relations of incompatibility and entailment among properties in terms of our grip on the functions with which properties are identified. Two properties are incompatible just in case, for any possible world, their extensions are disjoint. So, given the first world shown above, one's grasp of the property of being grey enables one to single out the set $\{a, b\}$ and one's grasp of the property of being white enables one to single out the set $\{c\}$, and, given the second world shown above, one's grasp of the property of being grey enables one to single out the set $\{a\}$ and one's grasp of the property of being white enables one to single out the set $\{b\}$. Clearly, these two sets are disjoint, and so it is for any possible world. As such, the two properties are incompatible. Our grasp of the incompatibility between these two properties is modeled as our grasp of two

functions such that, whenever they are given the same world as an input value, the sets of objects that are their output values are disjoint. However, if we can't model our grip on a property as our grip on a function from worlds to extensions, then we can't model our grip on the incompatibility of two properties in terms of our grip on two such functions. We can't, so the explanation of (F3) offered by our extra-worldly semantics goes out the window.

Now, as we've said, an extra-worldly semanticist will often not pay too much theoretical attention to the way in which the semantic theory assigns semantic value to content words, and so may, at this point, simply say that they are not concerned with explaining facts such as (F3). Consider, however, a set of facts for which an extra-worldly semanticist generally *will* want to provide an explanation. The sentence of "*a* is gray and *b* is white" entails the sentence "*a* is darker than *b*," but not vice versa. However, if you stick a negation operator in front of these two sentences, the entailment relation goes in the other direction: the sentence "It's not the case that *a* is darker than *b*" entails the sentence "It's not the case that (*a* is gray and *b* is white)," but not vice versa. One might take a possible worlds semantic framework to be able to explain this fact. In such a framework, atomic sentences are assigned sets of worlds as semantic values and "and" expresses the operation of taking the intersection of two sets of worlds. The set of worlds assigned to "*a* is gray" will be the set of worlds in which *a* is gray, the set of worlds assigned to "*b* is white" will be the set of worlds in which *b* is white, and the set of worlds assigned to "*a* is gray and *b* is white" will be the intersection of these two set of worlds. The set of worlds in which "*a* is darker than *b*," As a matter of fact, the first set is a subset of the second one. So, thinking of the entailment relation in terms of the subset relation, "*a* is gray and *b* is white" entails "*a* is darker than *b*." Insofar as there are some worlds in which *a* is darker than *b* but it is not the case that *a* is gray and *b* is white (for instance, any world in which *a* is black and *b* is white), the converse is not true. Now, by taking "It's not the

case that" to express complementation, it seems that we can explain why, when we embedded these sentences under this negation operator, the entailments are reversed. Complementation reverses the subset/superset relation between sets. If $A \subseteq B$ then $(B)' \subseteq (A)'$. So, by understanding entailment in terms of the subset relation and understanding negation in terms of complementation, we can explain why the entailment relations are reversed when the sentences are negated, and, by taking speakers' knowledge of meaning to be (adequately modeled by) knowledge of these semantic values, we can explain their knowledge of this fact. If I am right, however, there is no explanation to be had here.

The explanandum here—the fact to be explained—is that the entailment relations between sentences are reversed when these sentences are negated. The explanans—the set of facts doing the explaining—is that (1) sentences have sets of possible worlds as semantic values, (2) entailment between sentences is a matter of the subset relation obtaining between their semantic values, and (3) a negation operator semantically expresses the operation of complementation. I take the argument that I have just given to undermine (1), and, in doing so, undermined the whole explanation. Since the theory is compositional, such that the semantic value of a sentence must be composed of the semantic values of its parts, then, if functions from worlds to extensions cannot serve as semantic values of predicates, then sets of possible worlds cannot serve as the semantic values of sentences. If sets of possible worlds cannot serve as the semantic values of sentences, and so we cannot think of entailment in terms of the subset relation, nor can we think of negation in terms of complementation. If we can't think of entailment in terms of the subset relation and negation in terms of complementation, we have no explanation for the fact to be explained. So, if the argument I have just given goes through, then possible worlds semantics can only function as a tool for elucidating facts such as the fact that entailment relations between sentences are reversed when these sentences are negated. We can use possible worlds semantics as a

tool for systematically bringing into view the set of semantic relations that obtain between sentences of the language for which we are giving a semantic theory. In general, we can use model-theoretic tools to elucidate inferential relations, and possible worlds semantics for natural language is one instance of this general fact. However, we cannot, with the use of such a framework, explain what it is in virtue of which these semantic relations obtain. Since semantics, as I am understanding it here, is an explanatory enterprise, possible worlds semantics cannot be enough; it must be supplemented by a semantic framework that really is able to do this explanatory work.

2.8 Conclusion: The Limits of Possible Worlds Semantics

While possible worlds semantics may be a fine tool for systematizing a set of semantic facts of which we already have knowledge, we are not going to get an account of these facts or our knowledge of them through the use of a possible worlds semantics. Possible worlds semantics is not going to give us an account of linguistic meaning or our knowledge thereof. For such an account, we must look towards a different sort of semantic theory. In the next chapter, I'll consider an alternative sort of worldly semantic theory—intra-worldly semantics—focusing on the variant of such a semantic theory proposed by Scott Soames and Jeff King.