THE INCONSISTENCY THEORY OF TRUTH

John Barker

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
PHILOSOPHY

June 1998

Abstract

This dissertation uses the Liar paradox to motivate an account of the concept of truth that I call the "inconsistency theory of truth." The Liar paradox is the puzzle that arises when we consider such sentences, known as Liar sentences, that say of themselves that they are not true: whatever truth value we attribute to such a sentence, we seem to be immediately driven to the conclusion that it has the opposite truth value. Examining this puzzle reveals that its source is the following principle, which we may call the *disquotational* principle: for any sentence 'A', 'A' is true if and only if A. While this principle is very natural, it is also inconsistent, as is shown when the sentence 'A' is a Liar sentence.

After the introductory Chapter 1, Chapter 2 presents some of the technical work on the Liar. Chapter 3 discusses several accounts based on the idea that Liar sentences lack truth values; there I argue that those accounts fail to adequately address a problem known as the *strengthened Liar* problem. Next, Chapter 4 examines a family of views based on the idea that the formal contradictions derivable by means of the disquotational principle are not really genuine contradictions: according to these views, the appearance of contradiction results from ignoring a contextually determined hidden parameter. I argue that they, too, do not successfully handle the strengthened Liar problem.

In Chapter 5 I put forth a view based on the idea that the concept of truth itself is inconsistent. Namely, the view I defend is that the linguistic conventions governing the use of the word 'true' commit us to the disquotational schema which, as I said, is inconsistent. I conclude in Chapter 6 by considering some of the implications of believing the inconsistency theory.

Acknowledgements

I owe a pleasant debt of gratitude to my advisors, John Burgess and Scott Soames, for their help and encouragement at various stages in the writing of this dissertation. Without their helpful suggestions, useful criticisms and high standards, this dissertation certainly would have been much poorer.

I have been fortunate to have had the opportunity to talk to many other people about truth and the Liar, and what they had to say was generally quite helpful. Among those who have improved this dissertation and my thinking about truth in various ways are Paul Benacerraf, Jenann Ismael, Mark Johnston, Mark Kalderon, Saul Kripke, Stephen Menn, Ruth Michaels, Laurie Paul, Steven Quevedo, Gideon Rosen, Lionel Shapiro, Jennifer Saul, Jamie Tappenden, and Michael Thau.

Finally, I would like to thank audiences at Princeton University and the University of Pittsburgh, where portions of this dissertation were presented.

Contents

Abstract iii

Acknowledgements iv

Chapter 1. Introduction 1

Chapter 2. Technical Aspects of the Liar 7
1. Semantics 7
2. Proof Theory 24

Chapter 3. Truth Value Gap Accounts 33
1. Truth and Determinate Truth 34
2. McGee's Response 40

Chapter 4. The Hierarchy Approach 49
1. Introduction 49
2. Burge 60
3. Parsons 75
4. Barwise and Etchemendy 90
5. Gaifman 113
6. Kripke and Soames 118
Appendix: Burge's Construction 130

Chapter 5. The Inconsistency Theory 145
1. Where We Are 145
2. What Are Disquotational Properties? 147
3. What the Theory Is 150
4. What the Theory Isn't 157
5. Why Believe the Theory? 160
6. Objections Considered 165

Chapter 6. Conclusion 179

Bibliography 187

CHAPTER 1

Introduction

Understanding truth has long been impeded by a very old problem, which can be seen by considering the following sentence:

(1) (1) is false.

The difficulty is in assigning a consistent truth value to (1). Assume first that (1) is true; then since (1) says that (1) is false, (1) is false. Next assume that (1) is false. Since this is exactly what (1) says, (1) is true. So if (1) is true, then it's false; and if it's false, then it's true. How can this be?[1]

Sentence (1) is called a *Liar* sentence, and the problem just sketched is one version of the *Liar paradox*, the subject of this essay. That problem turns out to be a rather difficult one, more difficult perhaps than the foregoing suggests; I will eventually use it to motivate a theory of truth that may strike some as radical.

Strictly speaking, the conclusion reached one paragraph back is not a contradiction. It implies only that (1) is either both true and false, or neither true nor false. And while most (but not all) philosophers have considered the former unacceptable, many have been attracted to the latter. After all, there are plenty of clear examples of truth value gaps in natural language, and (1) in particular strikes many as contentless. But while this may consistently explain (1), it has a harder time accounting for other Liar sentences. In particular, consider the sentence

(2) (2) is not true.

[1]One might be tempted to say that (1) is malformed or involves some sort of illegitimate self-reference, and that this is what is causing the confusion. But this would be a mistake. '(1)' might well be viewed as an abbreviation for the expression, 'the first numbered sentence in John Barker's dissertation'; in that case, the fact that (1) refers to itself is clearly unproblematic.

This differs from (1) only in that 'false' has been replaced by 'not true'; but this seemingly small change defeats the truth-value-gap explanation of (2). For suppose (2) is gappy, i.e., neither true nor false; then in particular it is not true. But since what (2) *says* is that (2) is not true, it appears that (2) is true after all. Thus, apparently, the truth value gap account of the Liar saddles us with the conclusion that (2) is both true and not true; and this *is* a contradiction.

The Liar sentence (2) is called a *strengthened Liar*, and the problem suffered by the gap account is one instance of a general problem known as the *strengthened Liar problem*. This problem is really a strategy for defeating accounts of truth and the Liar: in short, the very terms by which the account is expressed are used to construct a new Liar sentence that the account is unable to handle. The strengthened Liar problem is a recurring theme both in the philosophical literature on the Liar and in this essay.

Let's look at another instance of the strengthened Liar problem. Thus far I have treated truth as a property belonging to sentences. This is fairly common in discussions of the Liar, even though it is somewhat artificial; and this has led many to suspect that the paradox would somehow go away if we took the more natural approach of treating truth as a property of propositions. Then, one might think, Liar sentences could simply be explained as sentences that fail to express propositions. The idea that Liar sentences express no propositions is independently appealing, since in considering the Liar one is rather at a loss as to just what it says. In particular, one might well think that a claim 'A is true' or 'A is not true' gets its content from that of 'A', in which case the Liar never gets any content at all: for (2) to have a content, the sentence to which nontruth is ascribed, which is simply (2) itself, would have to have a content *already*, which is absurd.[2]

[2]This is a very informal way of saying that (2) is ungrounded; we will give a more formal treatment of ungroundedness in the next chapter.

This strategy simply won't work, as another application of the strengthened Liar strategy shows. Consider the sentence

(3) (3) does not express a true proposition.

Let us call (3) a *propositional Liar sentence*. Claiming that (3) expresses a true proposition, or that it expresses a false proposition, leads to a contradiction in much the way assigning a truth value to (2) does, so presumably (3) expresses no proposition at all. Now, however, notice that if a sentence expresses no proposition, then *a fortiori* it expresses no *true* proposition. So we are now in the position of claiming that (3) does not express a true proposition; in so doing, we are assertively uttering a sentence—namely (3)—that we are also claiming says nothing. This is surely self-defeating.

The contradiction that we derived from the claim that (2) is gappy can apparently be derived without the help of that claim. Just as we showed that (1) is true just in case it's false, we can apparently show that (2) is true just in case it's not true; and this is itself contradictory, provided the usual rules of classical logic apply. (We will later consider reasons for doubting that the rules of classical logic apply here.) Namely: either (2) is true or it isn't; if it is, then it's also not true; if it's not true, then it's also true; so either way (2) is true and at the same time not true, a contradiction.

We seem to have derived a contradiction out of thin air. We shouldn't be able to do this. Therefore, we ought to take a closer look at the reasoning that generated the contradiction, to see exactly where it goes wrong; doing so will reveal why the Liar paradox is a problem in our understanding of truth, not just in our understanding of certain odd sentences. The reasoning we used to derive the contradiction can be reconstructed as follows:

1. (2) is true (Assumption)
2. '(2) is not true' is true (from 1)
3. (2) is not true (from 2)

4.	If (2) is true then (2) is not true	(from 1–3)
5.	(2) is not true	(Assumption)
6.	'(2) is not true' is true	(from 5)
7.	(2) is true	(from 6)
8.	If (2) is not true then (2) is true	(from 5–7)
9.	(2) is true or (2) is not true	(excluded middle)
10.	(2) is true and (2) is not true	(from 4, 8 and 9)

In going from 1 to 2, and from 6 to 7, we use the substitution of equals for equals, together with the premise

(4) $$(2) = \text{'(2) is not true'}$$

(4) seems unassailable.[3] The remaining inferences are all (classically) logically valid, with the exception of those between 2 and 3 and between 5 and 6. The former appears to use the general rule

$$\mathrm{T}_1: \quad \frac{\text{'}S\text{' is true}}{S}$$

and the latter appears to use

$$\mathrm{T}_2: \quad \frac{S}{\text{'}S\text{' is true}}$$

To avoid a contradiction, it is therefore necessary to reject either (4), or one of the rules T_1 and T_2, or part of classical logic. But it is very far from clear how we could do any of these things.

We can similarly derive a contradiction using (4), classical logic, and the schema

(T) $$\text{'}S\text{' is true iff } S$$

(The reasoning is essentially the same, so I will omit it.) Again, it is not obvious what it is open to us to reject here. If the problem lies with (T) (or with T_1 and T_2),

[3]Statements like (4) are most naturally viewed as empirical claims. Alternatively, using a trick due to Gödel, we can interpret them so that they turn out to be mathematical truths. The trick can be found in the proof of Theorem 1 in the next chapter.

then there is something lacking, to say the least, in our understanding of truth, since schema (T) seems absolutely basic to that understanding. Far from being a substantial philosophical assumption, (T) seems like a trifling definitional truth, so obvious that we weren't at first explicitly aware that we were even using it in our reasoning about the Liar. Even if there turns out to be nothing wrong with (T) and the problem lies wholly in our use of classical logic, there is still something lacking in our understanding of truth, since as yet we have not seen why classical logic should not apply to reasoning about Liar sentences.

Philosophers have made many attempts to find out where our naive understanding of truth goes wrong, and to correct that understanding. We will examine some of these in Chapters 3 and 4. These all proceed under the assumption that Liar reasoning (as I will call arguments like the foregoing) involves some sort of mistake, a false assumption or an invalid inference, that can be corrected by finding a more accurate theory of truth. We will see that the strengthened Liar problem is the bane of these theories. Whatever insights they provide, and whatever other defects they may have, they all have trouble dealing with one form or another of the strengthened Liar. This is so even for those theories that are specifically designed to *avoid* the strengthened Liar problem.

Moreover, one soon gets the sense in considering these theories that *any* theory of truth could be undermined by a suitable application of the strengthened Liar strategy. But if this is right, then we are left in an odd position. After all, one would think, *some* theory of truth must be right.

I am going to suggest an alternative approach, which *admits* that the strengthened Liar is unavoidable rather than trying to avoid it. On this view, there is no alternative theory of truth that is more accurate than our naive understanding of truth, even though the latter is inconsistent. The inconsistency is built into the very concept of truth itself. Specifically, schema (T) seems obvious to us not because it is a near approximation to the correct theory of truth, but because acceptance of (T) is all

there is to the fact of our meaning *true* by 'true'. Thus, there is a clear sense in which the *concept* of truth, not just this or that *theory* of truth, is inconsistent. For this reason, I call the view the "inconsistency theory of truth." This view is not entirely new by any means; the basic idea was stated in [**Tar35**], and developed more fully in [**Chi79**]. Numerous other philosophers have also found it appealing. But on the whole the view has received very little in the way of sustained articulation or defense.

The next chapter summarizes part of the extensive technical research on the Liar, including some proof-theoretic results that may not be well known among philosophers. This will provide a more precise account than I have given thus far as to what is and is not consistent to say about truth. Chapter 3 examines some truth value gap accounts of the Liar that are more sophisticated than the account sketched above, but which ultimately cannot deal adequately with the strengthened Liar problem. Chapter 4 considers several incarnations of what appears to be currently the most popular approach to the paradoxes, which I have labeled the "hierarchy approach." There I conclude that when nothing else is wrong with such theories, they ask us to accept a view that, by its own lights, cannot be stated; while some philosophers find this sort of thing appealing, those of us who don't will naturally find the hierarchy approach seriously flawed. Chapter 5 introduces the inconsistency theory of truth and defends it against several objections. Finally, the conclusion briefly discusses the upshot of believing the inconsistency theory, both practical and theoretical.

CHAPTER 2

Technical Aspects of the Liar

This chapter develops certain technical facts and concepts that will prove useful later on. The point is not to make any technically novel contributions, but to summarize the technical material that we will need later. A reader uninterested in technical details may safely skim this chapter and refer back to it as necessary.

1. Semantics

In studying the concept of truth in *natural* languages, philosophers have often found it useful to study it in connection with *formal* languages. This is not because there is a detailed similarity between the two. There isn't: the languages we will be studying here lack many of the features found in natural language, such as ambiguity, indexicality, tense, mood, non-referring terms, presupposition, sortal incorrectness and, at least in the case of classical languages, vagueness. This lack is precisely their strength. Formal languages are idealized models of natural languages; studying truth in connection with them allows us to ignore complicating irrelevances like the ones just mentioned. Of course, this strategy assumes that they *are* irrelevant, something we cannot know for sure in advance; but it seems reasonable to assume that they are until we find a compelling reason to think otherwise.

1.1. Truth In Classical Languages. The first rigorous definition of 'true sentence' for a formal language was that of Tarski's [**Tar35**]. For Tarski, the general problem was to give, within a formal language $\mathcal{M}$ (the "metalanguage"), a definition of 'true sentence of $\mathcal{L}$' for some other formal language $\mathcal{L}$ (the "object language"). More precisely, assume that the sentences of $\mathcal{L}$ can be translated into $\mathcal{M}$, and that $\mathcal{M}$ also has a "structural description" d_φ of each sentence φ of $\mathcal{L}$. (What exactly a

structural description is needn't concern us here.) Tarski's project was to define, in $\mathcal{M}$, an expression Tr such that for each sentence φ of $\mathcal{L}$,

(T) $$Tr(d_\varphi) \leftrightarrow \varphi'$$

is provable,[1] where φ' is φ's translation into $\mathcal{M}$. Any definition satisfying this condition is called a *materially adequate* definition of truth-in-$\mathcal{L}$. Now (T) is simply a formal version of the disquotational schema that got us into trouble in the first place. Tarski's discovery was that, while in general it is impossible to give in $\mathcal{M}$ a materially adequate definition of truth-in-$\mathcal{L}$ when $\mathcal{M} = \mathcal{L}$, it often *is* possible to do so when $\mathcal{M} \neq \mathcal{L}$. More generally, a materially adequate definition of truth will be impossible when every formula of $\mathcal{M}$ is translatable into $\mathcal{L}$ (except in the most trivial cases); for a definition of truth in the object language to be possible in the metalanguage, the latter must (usually) be *essentially richer* than the former.

What Tarski gave in [**Tar35**] was a general strategy for producing a definition in $\mathcal{M}$ of 'true sentence of $\mathcal{L}$' given any *particular* $\mathcal{M}$ and $\mathcal{L}$; a more modern version of this idea is the notion of *truth in a (classical) model.* (The notion of model is also due to Tarski.) The notions I will use here are quite standard, and I will confine myself to a summary. By a *language* we mean a set of non-logical constants (i.e., predicates, individual constants, and function symbols). Thus a language is an *uninterpreted* language; this use of 'language' is somewhat unfortunate, and contrasts with that of the last paragraph, but it is also standard. A *model* for $\mathcal{L}$ is a pair $\mathfrak{M} = (M, f)$, where M is a nonempty set (called $\mathfrak{M}$'s *domain*), and f assigns appropriate objects to the symbols in $\mathcal{L}$. The *formulas* of $\mathcal{L}$ are defined in the standard way; I will take the logical primitives to be $\neg$, $\vee$ and $\exists$.

Given a model $\mathfrak{M}$ that interprets a language $\mathcal{L}$, let $\mathcal{L}_M$ be the language obtained from $\mathcal{L}$ by adding a new constant $\overline{x}$ for each $x \in M$ (with distinct $\overline{x}$'s for distinct

[1]Tarski thought of languages as containing certain axioms and inference rules, and 'provable' here means provable on the basis of them. We now locate these axioms and rules in theories, rather than languages.

x's); truth in $\mathfrak{M}$ of sentences of $\mathcal{L}$ is defined by first defining the more general notion of truth in $\mathfrak{M}$ of a sentence of $\mathcal{L}_M$. For each term t of $\mathcal{L}_M$, let $t^{\mathfrak{M}}$ be t's denotation in $\mathfrak{M}$; let us assume that denotation has been defined in some standard way. The *explicit* definition of truth in $\mathfrak{M}$ is then based on the following *inductive* definition:

(1) If P is an n-place predicate of $\mathcal{L}$ that $\mathfrak{M}$ interprets as the relation R and if $t_1 \dots t_n$ are terms of $\mathcal{L}_M$, then $P(t_1 \dots t_n)$ is true in $\mathfrak{M}$ iff $(t_1^{\mathfrak{M}} \dots t_n^{\mathfrak{M}}) \in R$;

(2) $\neg\varphi$ is true in $\mathfrak{M}$ iff φ is not true in $\mathfrak{M}$;

(3) $(\varphi \vee \psi)$ is true in $\mathfrak{M}$ iff φ is true in $\mathfrak{M}$ or ψ is true in $\mathfrak{M}$ (or both);

(4) $\exists x\, \varphi$ is true in $\mathfrak{M}$ iff there is an a in $\mathfrak{M}$'s domain such that $\varphi(\overline{a})$ is true in $\mathfrak{M}$ (where $\varphi(\overline{a})$ is the result of replacing the free occurrences of x in φ by $\overline{a}$).

As a definition this leaves something to be desired, since the *definiendum* appears in the *definiens*. But this circle is not a vicious one: we can prove (in our background mathematical theory) that for any $\mathcal{L}$ and any model $\mathfrak{M}$ for $\mathcal{L}$, there is a unique set Tr of sentences of $\mathcal{L}_M$ that satisfies (1)–(4) (i.e., such that (1)–(4) hold when 'is true in $\mathfrak{M}$' is replaced by 'belongs to Tr'). A sentence is now defined to be true in $\mathfrak{M}$ just in case it belongs to this set.

Our definition of 'true in $\mathfrak{M}$' was given in an informal metalanguage, and we were not terribly specific about what background assumptions are needed to prove the existence and uniqueness of the set Tr. Proceeding more formally, we could take the metalanguage to be the language of set theory, and our background theory to be ZFC (say). Letting $\Phi(\mathfrak{M}, y)$ be a formalization of (1)–(4) with 'is true in $\mathfrak{M}$' replaced by '$\in y$', we can now prove in our background set theory that for each $\mathfrak{M}$, there is a unique y such that $\Phi(\mathfrak{M}, y)$. Either $\exists y\, (\Phi(\mathfrak{M}, y) \wedge x \in y)$ or $\forall y\, (\Phi(\mathfrak{M}, y) \rightarrow x \in y)$ will serve as a definition of 'x is a true sentence of $\mathfrak{M}$' in the language of set theory. Given a reasonable way of translating sentences of $\mathcal{L}$ into the language of set theory, each instance of (T) is provable (though we will not prove this here).

An n-tuple $(a_1 \dots a_n)$ from M is said to *satisfy* a formula $\varphi(x_1 \dots x_n)$ of $\mathcal{L}_M$ in $\mathfrak{M}$ just in case the sentence $\varphi(\overline{a_1} \dots \overline{a_n})$ is true in $\mathfrak{M}$. If we had not introduced the extra

constants of $\mathcal{L}_M$, confining ourselves throughout to formulas of $\mathcal{L}$, then we would have to have *first* defined satisfaction in $\mathfrak{M}$, and *then* defined truth in $\mathfrak{M}$ in terms of satisfaction; the point of the present approach is to avoid this detour.

Now that we have seen how to give a Tarski-style definition of truth-in-a-model, let's examine the result mentioned briefly before, that in general it is not possible to give, in a given language, a definition of truth for that language. To make this precise, we need to define definability. If $\varphi = \varphi(x_1 \dots x_n)$ is a formula of $\mathcal{L}_M$ with all free variables displayed, the relation $R = \{(a_1 \dots a_n) : \varphi(\overline{a_1} \dots \overline{a_n})$ is true in $\mathfrak{M}\}$ is called the relation *defined* by φ in $\mathfrak{M}$. A relation is *definable* in $\mathfrak{M}$ if it is the relation defined in $\mathfrak{M}$ by some formula of $\mathcal{L}_M$; a relation is *definable* in $\mathfrak{M}$ *without parameters* if it is defined in $\mathfrak{M}$ by some formula of $\mathcal{L}$. ('Definable *set*' is defined similarly.) Let us now assume that the sentences of $\mathcal{L}_M$ are elements of $\mathfrak{M}$'s domain, and instead of $\overline{\varphi}$ let us write 'φ'. Our question then becomes: is the set of true sentences of $\mathcal{L}_M$ (or of $\mathcal{L}$) definable in $\mathfrak{M}$?

Then answer is *no*, provided $\mathfrak{M}$ is capable of expressing certain basic syntactic notions; specifically, truth is indefinable in $\mathfrak{M}$ provided the *substitution relation* is definable in $\mathfrak{M}$, where the substitution relation is $\{(\varphi(x), a, \varphi(\overline{a})) : \varphi(x)$ a formula of $\mathcal{L}_M$ and $a \in M\}$. In that case we can apply the following:

THEOREM 1 (Gödel's self-reference lemma). *Assume substitution is definable in* $\mathfrak{M}$; *then for each formula* $\varphi(x)$ *of* $\mathcal{L}_M$ *of one free variable, there is a sentence* γ *of* $\mathcal{L}_M$ *such that* $\gamma \leftrightarrow \varphi(\text{`}\gamma\text{'})$ *is true in* $\mathfrak{M}$.

PROOF. Let $\sigma(x, y, z)$ define the substitution relation in $\mathfrak{M}$, let $\psi(x)$ be the formula $\exists y\,(\sigma(x, x, y) \wedge \varphi(y))$, and let γ be the sentence $\psi(\text{`}\psi(x)\text{'})$. Then $\sigma(\text{`}\psi(x)\text{'}, \text{`}\psi(x)\text{'}, y)$ says that $y = \psi(\text{`}\psi(x)\text{'})$, so γ, which is the sentence $\exists y\,(\sigma(\text{`}\psi(x)\text{'}, \text{`}\psi(x)\text{'}, y) \wedge \varphi(y))$, is equivalent to $\varphi(\text{`}\psi(\text{`}\psi(x)\text{'})\text{'})$, which is simply the sentence $\varphi(\text{`}\gamma\text{'})$. □

When $\gamma \leftrightarrow \varphi(\text{`}\gamma\text{'})$ is true in $\mathfrak{M}$, we will say that $\gamma$ *says* $\varphi(\text{`}\gamma\text{'})$ (following the convention of [**Bur86**]). From the self-reference lemma, the following is almost immediate.

THEOREM 2 (Tarski). *If substitution is definable in $\mathfrak{M}$, then the set of sentences of $\mathcal{L}_M$ that are true in $\mathfrak{M}$ is not definable in $\mathfrak{M}$.*

PROOF. Suppose the contrary, and let $\tau(x)$ be a formula of $\mathcal{L}_M$ that defines the set of truths of $\mathfrak{M}$. Let λ be a sentence of $\mathcal{L}_M$ that says $\neg\tau(\text{`}\lambda\text{'})$. Then $\lambda \leftrightarrow \neg\tau(\text{`}\lambda\text{'})$ is true in $\mathfrak{M}$; but since τ defines truth, $\lambda \leftrightarrow \tau(\text{`}\lambda\text{'})$ is also true in $\mathfrak{M}$, so $\lambda \leftrightarrow \neg\lambda$ is true in $\mathfrak{M}$, which is plainly impossible. □

REMARK. Examining this proof reveals that if the substitution relation is definable in $\mathfrak{M}$ without parameters, then the set of true sentences of the language $\mathcal{L}$ (as opposed to $\mathcal{L}_M$) is not definable in $\mathfrak{M}$ without parameters. It should also be noted that appropriate versions of these results still hold if we assume not that M literally contains the sentences of $\mathcal{L}_M$, but that it contains *codes* for these sentences. (For example, it is enough to assume that $\mathfrak{M}$ is *acceptable* in the sense of [**Mos74**].) From now on, we will assume that substitution is definable in $\mathfrak{M}$.

There is an obvious similarity between the above proof and the Liar paradox. Notice, however, that instead of resulting in a contradiction, our proof results in a theorem. Specifically, it results in an *indefinability* theorem; in fact, different versions of the Liar paradox can be parlayed into different indefinability results in just this way.

Before moving on, it may be worth mentioning that a certain related result can be obtained without assuming that substitution is definable in $\mathfrak{M}$.[2] Let $\mathfrak{M}$ be any infinite model that interprets a countable language, and let * be a 1-1 map from {formulas of $\mathcal{L}_M$ with at most one free variable} into $\mathfrak{M}$'s domain. (Think of φ^* as a code for φ.)

[2]The material in this paragraph is due to Saul Kripke.

* might be the inclusion map (in case $\mathfrak{M}$ actually contains the formulas of $\mathcal{L}_M$), or it might be anything else; in particular, it need not be "effective" in any sense. Relative to $\mathfrak{M}$ and *, define the satisfaction relation to be the relation $\{(\varphi^*, a) : \varphi$ belongs to *'s domain and $a \in M$ and $\varphi(\overline{a})$ is true in $\mathfrak{M}\}$. Then we have the following:

THEOREM 3. *For any $\mathfrak{M}$ and any * as specified above, the satisfaction relation (relative to $\mathfrak{M}$ and *) is not definable in $\mathfrak{M}$.*

PROOF. Suppose to the contrary that $\sigma(x, y)$ defined it. Then, letting $\eta = \eta(x)$ be the formula $\neg\sigma(x, x)$, we see that $\eta(\overline{\eta^*})$ is true in $\mathfrak{M}$ iff $\neg\sigma(\overline{\eta^*}, \overline{\eta^*})$ is true in $\mathfrak{M}$ iff $\eta(\overline{\eta^*})$ is not true in $\mathfrak{M}$, a contradiction. □

(As with Tarski's theorem, if we replace '$\mathcal{L}_M$' with '$\mathcal{L}$' in the definition of the satisfaction relation, the resulting relation is not definable without parameters.) In fact, Tarski's theorem follows directly from Theorem 3, since if substitution and truth were both definable in $\mathfrak{M}$ by formulas $Sub(x, y, z)$ and $Tr(x)$, then satisfaction would be defined in $\mathfrak{M}$ by $\exists z\,(Sub(x, y, z) \wedge Tr(z))$.[3]

1.2. Truth In Partial Languages. We have seen that a common idea about the Liar is that it is neither true nor false; since there are no truth value gaps in classical languages, the question naturally arises whether we can get around Tarski's indefinability theorem by allowing such gaps. It turns out that we can, as was demonstrated by Kripke in his groundbreaking [**Kri75**] (and independently by Martin and Woodruff in [**MW75**], which proves a special case of Kripke's main result).

To understand Kripke's result, we first need to say what we mean by a language that allows truth value gaps. If $\mathcal{L}$ is a set of nonlogical symbols, then a *partial model* for $\mathcal{L}$ is just like a classical model for $\mathcal{L}$, except that it assigns to each n-place predicate a disjoint pair of n-place relations on its domain; the first relation in the pair is called the predicate's *extension*, the second its *antiextension*. (Constants and function

[3]In [**Gup82**], Gupta makes much of the existence of classical languages that have their own truth predicates but not their own substitution predicates. As Theorem 3 shows, however, the existence of such languages is of limited importance, since *no* classical language has its own satisfaction predicate.

symbols are treated exactly as they are in classical logic.) The idea is that a predicate is true of the n-tuples in its extension, false of those in its antiextension, and undefined for the rest. If $\mathfrak{M}$ and $\mathfrak{M}'$ are partial models for the same language, we say that $\mathfrak{M}'$ *extends* $\mathfrak{M}$ if (1) they have the same domain and agree on the interpretations of all individual constants and function symbols, and (2) the extension (resp. antiextension) in $\mathfrak{M}$ of each predicate they interpret is a subset of its extension (antiextension) in $\mathfrak{M}'$. In other words, $\mathfrak{M}'$ differs from $\mathfrak{M}$ only in classifying objects that $\mathfrak{M}$ leaves unclassified; it doesn't *re*-classify objects. Similarly, if $\mathfrak{M}$ is a partial model for a language $\mathcal{L}$ and $\mathfrak{N}$ is a classical model for that same language, we say that $\mathfrak{N}$ is an extension of $\mathfrak{M}$ just in case (1) they have the same domain and agree on the interpretations of the individual constants and function symbols, and (2) for each predicate P, P's extension in $\mathfrak{N}$ contains its extension in $\mathfrak{M}$ and is disjoint from its antiextension in $\mathfrak{M}$.

A classical or partial model $\mathfrak{M}'$ for a language $\mathcal{L}'$ is said to be an *expansion* of a model $\mathfrak{M}$ for a language $\mathcal{L}$ if $\mathcal{L} \subseteq \mathcal{L}'$ and $\mathfrak{M}$ and $\mathfrak{M}'$ have the same domain and agree on the interpretation of the symbols in $\mathcal{L}$. In this case $\mathfrak{M}$ is said to be the $\mathcal{L}$-*reduct* of $\mathfrak{M}'$. The notion of expansion should not be confused with that of extension. If T is a unary predicate that does not belong to $\mathcal{L}$ and $\mathfrak{M}$ is a classical model for $\mathcal{L}$, then $(\mathfrak{M}, X)$ is defined to be the expansion of $\mathfrak{M}$ to $\mathcal{L} \cup \{T\}$ that interprets T as X. Likewise, if $\mathfrak{M}$ is a partial model for $\mathcal{L}$, $(\mathfrak{M}, (E, A))$ is the expansion of $\mathfrak{M}$ to $\mathcal{L} \cup \{T\}$ that interprets T as (E, A).

Defining truth in a partial model is less straightforward than it is for classical models. To see why, suppose a sentence φ is undefined in $\mathfrak{M}$ but that ψ is true in $\mathfrak{M}$. Should $\varphi \vee \psi$ be true in $\mathfrak{M}$, or merely undefined? Intuitions differ, and the answer probably depends on how we interpret the truth value gaps. Likewise, should $\varphi \vee \neg\varphi$ be undefined, or should it be true? Again, intuitions differ. Instead of trying to settle this sort of dispute in advance, let's adopt a framework that can accommodate as many intuitions as possible. A definition of truth and falsehood for partial models is

called a *valuation scheme*. There is no one standard definition of this term; I will use the following, which is very inclusive yet specific enough for our purposes.

DEFINITION. A *valuation scheme* is a pair $\sigma = (\models_\sigma, \dashv_\sigma)$ of relations that relate partial models to the formulas they interpret, such that

(1) We never have both $\mathfrak{M} \models_\sigma \varphi$ and $\mathfrak{M} \dashv_\sigma \varphi$;

(2) For any n-place predicate P and any $a_1 \dots a_n \in M$, $\mathfrak{M} \models_\sigma P(\overline{a_1} \dots \overline{a_n})$ iff $(a_1 \dots a_n) \in P$'s extension (in $\mathfrak{M}$) and $\mathfrak{M} \dashv_\sigma P(\overline{a_1} \dots \overline{a_n})$ iff $(a_1 \dots a_n) \in P$'s antiextension; and

(3) For any terms s and t, and any formula $\varphi(x)$ with at most x free, if $s^{\mathfrak{M}} = t^{\mathfrak{M}}$ then $\mathfrak{M} \models_\sigma \varphi(s)$ iff $\mathfrak{M} \models_\sigma \varphi(t)$ and $\mathfrak{M} \dashv_\sigma \varphi(s)$ iff $\mathfrak{M} \dashv_\sigma \varphi(t)$.

When $\mathfrak{M} \models_\sigma \varphi$ (resp. $\mathfrak{M} \dashv_\sigma \varphi$), we say that φ is *true* (*false*) in $\mathfrak{M}$ relative to σ; a sentence that is neither true nor false in $\mathfrak{M}$ relative to σ is *undefined* in $\mathfrak{M}$ relative to σ. There are many valuation schemes, but three have been more thoroughly studied than the others. The first is called the *strong Kleene* scheme, which we will denote by SK; it is the unique valuation scheme that satisfies the following:

- $\mathfrak{M} \models_{\text{SK}} \neg\varphi$ iff $\mathfrak{M} \dashv_{\text{SK}} \varphi$;
  $\mathfrak{M} \dashv_{\text{SK}} \neg\varphi$ iff $\mathfrak{M} \models_{\text{SK}} \varphi$.
- $\mathfrak{M} \models_{\text{SK}} (\varphi \vee \psi)$ iff $\mathfrak{M} \models_{\text{SK}} \varphi$ or $\mathfrak{M} \models_{\text{SK}} \psi$;
  $\mathfrak{M} \dashv_{\text{SK}} (\varphi \vee \psi)$ iff $\mathfrak{M} \dashv_{\text{SK}} \varphi$ and $\mathfrak{M} \dashv_{\text{SK}} \psi$.
- $\mathfrak{M} \models_{\text{SK}} \exists x\, \varphi(x)$ iff $\mathfrak{M} \models_{\text{SK}} \varphi(\overline{a})$ for some $a \in M$;
  $\mathfrak{M} \dashv_{\text{SK}} \exists x\, \varphi(x)$ iff $\mathfrak{M} \dashv_{\text{SK}} \varphi(\overline{a})$ for all $a \in M$.

The second scheme, the *weak Kleene* scheme (denoted WK), is just like the strong Kleene except that in each of the above biconditionals, the sentence mentioned on the left is undefined whenever any of the sentences mentioned on the right are undefined. (Thus a disjunction is undefined whenever one of its disjuncts is, even if its other disjunct is true.) We will not deal much with this scheme.

The third scheme is van Fraassen's *supervaluational* scheme, denoted vF. Truth and falsity in vF are defined as follows: $\mathfrak{M} \models_{\text{vF}} \varphi$ (resp., $\mathfrak{M} \models\!\!\!|_{\text{vF}} \varphi$) iff φ is true (false) in all classical models that extend $\mathfrak{M}$. Obviously, every classically valid sentence is true in any partial model relative to vF; in particular, any sentence $\varphi \vee \neg\varphi$ is true even when φ is undefined.

Our definition of "valuation scheme" lets in quite a lot, including some pretty strange beasts. For example, it's easy to find a valuation scheme that counts a non-atomic sentence true just in case the number of negation signs it contains is a Mersenne prime.[4] A decent theory of valuation schemes would certainly require a better definition. But my aim is not to develop a decent theory of valuation schemes; it is simply to provide a framework within which results about the definability and indefinability of truth can be stated and proven with a very high degree of generality.[5]

The schemes SK, WK and vF all have an important property, called *monotonicity*. A scheme σ is *monotone* iff whenever $\mathfrak{N}$ extends $\mathfrak{M}$ and φ is a sentence they both interpret, $\mathfrak{M} \models_\sigma \varphi$ implies $\mathfrak{N} \models_\sigma \varphi$ and $\mathfrak{M} \models\!\!\!|_\sigma \varphi$ implies $\mathfrak{N} \models\!\!\!|_\sigma \varphi$. (This is immediate for the scheme vF, and is shown by a routine induction on the complexity of sentences for the other two.) Monotonicity is the key to obtaining partial models with their own truth predicates.

We are almost ready to prove the basic results about the definability of truth in partial models. Similarly to the classical case, we may say that a formula φ of $\mathcal{L}$ *defines* a set X in $\mathfrak{M}$ (relative to σ) just in case $X = \{a : \mathfrak{M} \models_\sigma \varphi(\overline{a})\}$—notice that we do *not* require $\varphi(\overline{a})$ to be false when $a \notin X$, but merely to be not true. Thus, the set of truths of $\mathfrak{M}$ will be definable in $\mathfrak{M}$ provided there is a formula $\tau(x)$, satisfied

[4]Clause 2 of the definition settles the truth values of atomic sentences.

[5]Gupta and Belnap criticize Kripke in [**GB93**] for not defining 'valuation scheme', and they provide a definition of their own which is much narrower than the one given here. However, it should be clear that all Kripke means by 'valuation scheme' is a general definition of truth and falsehood in a partial model, and that he intends this notion to be a fairly broad one, so that his results will be as general as possible.

only by sentences of $\mathcal{L}_M$, such that for all such sentences φ,

$$\mathfrak{M} \models_\sigma \tau(\text{`}\varphi\text{'}) \quad \text{iff} \quad \mathfrak{M} \models_\sigma \varphi.$$

Let us say that a formula $\tau(x)$ of $\mathcal{L}_M$ is a *truth predicate* for $\mathfrak{M}$ if it satisfies the above and, in addition,

$$\mathfrak{M} \models\!\!\!|_\sigma \tau(\text{`}\varphi\text{'}) \quad \text{iff} \quad \mathfrak{M} \models\!\!\!|_\sigma \varphi.$$

Now let $\mathcal{L}$ be a language and let T be a unary predicate not occurring in $\mathcal{L}$. Kripke showed that for any monotone scheme σ and any $\mathfrak{M}$, there is a pair (E, A) such that $T(x)$ is a truth predicate for $(\mathfrak{M}, (E, A))$ relative to σ. By clause 2 of the definition of 'valuation scheme', this will hold provided $E = \{\text{true sentences of } (\mathfrak{M}, (E, A))\}$ and $A \cap \{\text{sentences}\} = \{\text{false sentences of } (\mathfrak{M}, (E, A))\}$.[6]

Kripke constructs such a pair by stages. Initially the extension and antiextension of the predicate T are empty. If T is interpreted as (E, A) at a given stage, then it is interpreted as (E', A') at the next stage, where $E' = \{\text{true sentences of } (\mathfrak{M}, (E, A))\}$ and $A' = \{\text{nonsentences}\} \cup \{\text{false sentences of } (\mathfrak{M}, (E, A))\}$. More formally, define $J_\sigma(E, A) = (E', A')$; then Kripke's construction yields the a sequence (E_0, A_0), (E_1, A_1), $\ldots (E_\alpha, A_\alpha), \ldots$ defined by

$$(E_0, A_0) = (\emptyset, \emptyset)$$

$$(E_{\alpha+1}, A_{\alpha+1}) = J_\sigma(E_\alpha, A_\alpha)$$

$$(E_\lambda, A_\lambda) = \left(\bigcup_{\xi<\lambda} E_\xi, \bigcup_{\xi<\lambda} A_\xi \right), \quad \text{limit } \lambda$$

Kripke showed that this sequence eventually reaches a limit, i.e., there is an α with $(E_\alpha, A_\alpha) = (E_\xi, A_\xi)$ for all $\xi > \alpha$. In particular, $(E_\alpha, A_\alpha) = (E_{\alpha+1}, A_{\alpha+1}) =$

[6]There is a tendency in the Liar literature to (mistakenly) credit Kripke only with proving the special case where σ is the scheme SK. The reason is that Kripke first proves his results for SK, then notes that the proofs actually work for any monotone scheme. Even as recently as [**GB93**], Gupta and Belnap do not give him credit for proving it for the scheme vF, even though he treats that scheme explicitly.

$J_\sigma(E_\alpha, A_\alpha)$, which is just to say that E_α and A_α are the sets of true and false sentences of $(\mathfrak{M}, (E_\alpha, A_\alpha))$, respectively (ignoring any nonsentences in M), which, as we have seen, implies that $(\mathfrak{M}, (E_\alpha, A_\alpha))$ has its own truth predicate.

Kripke's proof generalizes easily to a purely combinatorial result; rather than give Kripke's proof in its original form, we will derive his result from this more general fact. A *partially ordered set* is a pair $(X, \leq)$, where X is a set and $\leq$ is a reflexive, transitive, antisymmetric relation on X. (It is important that X be a set, not a proper class.) A subset Y of X is said to be *consistent* if every finite subset F of Y has an upper bound in X (i.e., there is an $x \in X$ such that $y \leq x$ for all $y \in F$). A partially ordered set X is a *complete coherent partial order* (ccpo) just in case each of X's consistent subsets Y has a least upper bound in X, which we denote $\bigvee Y$. For example, let X be the set of disjoint pairs of subsets of M, and define $\leq$ on X by $(E, A) \leq (E_0, A_0)$ iff $E \subseteq E_0$ and $A \subseteq A_0$; then X is a ccpo.

A function $f\colon X \to X$ is *monotone* if $x \leq y$ implies $f(x) \leq f(y)$. For example, if X is the set of disjoint pairs from M with the ordering of the last paragraph, then clearly the function J_σ is monotone just in case σ is a monotone valuation scheme. An $x \in X$ is a *sound point* of f if $x \leq f(x)$, and a *fixed point* if $x = f(x)$. Thus, a pair (E, A) is a J_σ-fixed point just in case $T(x)$ is a truth predicate in $(\mathfrak{M}, (E, A))$.

LEMMA 1. *If f is monotone and x is a sound point of f, then so is $f(x)$.*

PROOF. $x \leq f(x)$, so $f(x) \leq f(f(x))$ by monotonicity, which is just to say that $f(x)$ is sound. □

LEMMA 2. *If Y is a consistent set of sound points (relative to f), then $\bigvee Y$ is sound.*

PROOF. If $x \in Y$, then $x \leq \bigvee Y$ by definition, so $f(x) \leq f(\bigvee Y)$ by monotonicity; and $x \leq f(x)$ since x is sound, so $x \leq f(\bigvee Y)$. Since x was arbitrary, $f(\bigvee Y)$ is an upper bound for Y; and since $\bigvee Y$ is by definition Y's *least* upper bound, $\bigvee Y \leq f(\bigvee Y)$, i.e., $\bigvee Y$ is sound. □

THEOREM 4. *If f is a monotone function on a ccpo and x is a sound point of f, then f has a fixed point x' with $x \leq x'$; moreover, x' is the least such fixed point, i.e., if $x \leq y$ and y is fixed, then $x' \leq y$.*

PROOF. Put $x_0 = x$, and for each ordinal α define $x_{\alpha+1} = f(x_\alpha)$, $x_\lambda = \bigvee_{\xi<\lambda} x_\xi$ for limit λ provided $\bigvee_{\xi<\lambda} x_\xi$ exists (otherwise x_λ is undefined, as is each x_α for $\alpha > \lambda$). By induction on α we show that each x_α exists and is sound. Obviously x_0 is sound, and by Lemma 1, $x_{\alpha+1}$ is sound if x_α is. If λ is a limit and x_ξ is defined and sound for $\xi < \lambda$, then $x_\xi \leq x_{\xi+1}$ for all such ξ, and so $x_\xi \leq x_\zeta$ for all $\xi < \zeta < \lambda$. Thus $\{x_\xi : \xi < \lambda\}$ is linearly ordered by $\leq$, so it is certainly consistent; so $x_\lambda = \bigvee_{\xi<\lambda} x_\xi$ exists, and by Lemma 2 it is sound.

So x_α exists and is sound for all α, and $\alpha < \beta$ implies $x_\alpha \leq x_\beta$. Now either $x_\alpha < x_{\alpha+1}$ for all α, or $x_\alpha = x_{\alpha+1}$ for some α. In the former case the x_α are all distinct, so $\{x_\alpha : \alpha \in \mathrm{ON}\}$ is in one-to-one correspondence with ON (where ON is the class of ordinals); but that is impossible, since ON is a proper class and X is a set. So $x_\alpha = x_{\alpha+1} = f(x_\alpha)$ for some α, i.e., x_α is a fixed point. Also, $x = x_0 \leq x_\alpha$ since $0 \leq \alpha$. Finally, to see that x_α is the *least* fixed point extending x, it suffices to show that if y is a fixed point and $x \leq y$, then $x_\xi \leq y$ for all ξ; this is shown by a routine induction on ξ. □

In particular, since the set $\emptyset$ is trivially consistent, every ccpo has a $\leq$-least element $0 = \bigvee \emptyset$, and 0 is trivially sound; taking $x = 0$ in Theorem 4, we see that each monotone function on a ccpo has a least fixed point.[7]

Returning to partial models, we have already seen that J_σ is a monotone function on the set of disjoint pairs of subsets of M (provided σ is a monotone scheme), so it has a least fixed point (E, A); and we have also seen that $T(x)$ is a truth predicate for $(\mathfrak{M}, (E, A))$ just in case (E, A) is a fixed point. More generally, whenever (E, A) is a sound point, there is a fixed point $(E', A') \geq (E, A)$ (and indeed a least one).

[7]For an extended treatment of generalizations of Kripke's results to ccpo's, see [**Vis89**].

(E, A) is sound just in case every sentence in E (respectively, in A) is true (false) in $(\mathfrak{M}, (E, A))$; so Theorem 4 entails that whenever the predicate T classifies as true or false only those sentences that actually are true or false in $(\mathfrak{M}, (E, A))$, then its extension and antiextension may be enlarged in such a way that $T(x)$ is a truth predicate in the resulting partial model.

Since the existence of fixed point models depends only on the very abstract Theorem 4, we should expect that Kripke's results hold for a broader class of formalized languages than we have so far considered. This is indeed the case. One way to extend these results is to admit new logical vocabulary. We have so far assumed that all the formulas of a given language are built up from the constants of that language, the individual variables, and the logical symbols $\neg$, $\vee$ and $\exists$, but we could add new connectives and/or quantifiers. Kripke's result on the existence of fixed points will hold so long as this is done in such a way that truth and falsehood remain monotone, i.e., so long as the analogue of J_σ is a monotone function. For example, we might add a new quantifier $\exists^\infty$ ("there exist infinitely many"), and for our definition of $\models$ and $=\!\!|$ take the clauses in the definition for the scheme SK, plus the clauses

$$\mathfrak{M} \models \exists^\infty x\, \varphi(x) \quad \text{iff} \quad \{x : \mathfrak{M} \models \varphi(\overline{x})\} \text{ is infinite}$$

$$\mathfrak{M} =\!\!| \exists^\infty x\, \varphi(x) \quad \text{iff} \quad \{x : \mathfrak{M} =\!\!| \varphi(\overline{x})\} \text{ is cofinite}$$

One way of enlarging the set of logical symbols that does *not* result in a monotone definition of truth and falsehood is this: add a new unary connective $\sim$, and define it via the clauses

$$\mathfrak{M} \models \sim\varphi \quad \text{iff} \quad \text{not } \mathfrak{M} \models \varphi$$

$$\mathfrak{M} =\!\!| \sim\varphi \quad \text{iff} \quad \text{not } \mathfrak{M} =\!\!| \varphi$$

(the clauses for the other logical symbols being the same as with SK). A connective with this truth definition is called *external negation*, as opposed to the *internal* negation sign $\neg$. This truth definition is not monotone, since if φ is undefined in $\mathfrak{M}$ but true in an extension $\mathfrak{M}'$ of $\mathfrak{M}$, then $\sim\varphi$ is true in $\mathfrak{M}$ but false in $\mathfrak{M}'$. And indeed there are no fixed points with respect to this truth definition, as is seen by considering a sentence φ that says $\sim T(\text{`}\varphi\text{'})$. (See below for when we may assume that there is such a φ.)

Another way of generalizing Kripke's result is by changing our definition of 'partial model'. This might be accompanied by new logical symbols, e.g., modal or tense operators. Again, provided the function corresponding to J_σ is monotone, fixed points will exist.

Some natural generalizations actually require us to *restrict* the class of partial models. In particular, Kripke considers several ways of modifying the supervaluational scheme vF that require such a restriction. On one such modification, we count a sentence true (false) in $(\mathfrak{M}, (E, A))$ just in case it is true (false) in each of those of its classical extensions $(\mathfrak{M}, U)$ where U is a first-order consistent set of sentences. Another is just the same except that we require U to be *maximal* consistent. On either scheme we run into a bit of difficulty when we try to evaluate a sentence for truth or falsity in a model $(\mathfrak{M}, (E, A))$ for which no such consistent U exists. Taken literally, the definition of truth just given would require each sentence to be both true and false in such a model. We might try to avoid this by stipulating that, say, every sentence is undefined in such a model. But however we handle such models, the function J_σ will turn out to be a monotone function not on the set of all disjoint pairs, but merely on the set of disjoint (E, A) for which $E^* = E \cup \{\varphi : \neg\varphi \in A\}$ is consistent. (To see this, suppose J_σ were monotone on the set of all disjoint pairs. Then given a consistent sentence φ whose negation is also consistent, $T(\text{`}\varphi\text{'})$ is true in $(\mathfrak{M}, (\{\varphi\}, \emptyset))$ and false in $(\mathfrak{M}, (\{\neg\varphi\}, \emptyset))$; by monotonicity, $T(\text{`}\varphi\text{'})$ is then both true

and false in $(\mathfrak{M}, (\{\varphi, \neg\varphi\}, \emptyset))$, which is impossible.) Therefore, to apply Theorem 4 we must verify that the set of disjoint (E, A) for which E^* is consistent is a ccpo.

It turns out that it is; the proof uses the compactness of first-order logic. Indeed, if L is any logic, consider the family $\mathcal{P}(L)$ of disjoint pairs (E, A) such that E^* is L-consistent; it is easy to see that $\mathcal{P}(L)$ is a ccpo if and only if L is compact. Let σ be the valuation scheme such that for L-consistent E^*, φ is true (false) in $(\mathfrak{M}, (E, A))$ relative to σ just in case φ is true (false) in $(\mathfrak{M}, U)$ for all L-consistent U containing E and disjoint from A. Then for non-compact logics, Theorem 4 does not automatically imply that J_σ has any fixed points, even though it is obviously monotone. Indeed, McGee has found a logic for which σ has no fixed points.[8]

What about Liar sentences, by which we showed that no classical model has its own truth predicate? These simply turn out to lack truth values. Similarly to the last section, let us say that a sentence φ *says* $\psi(\text{`}\varphi\text{'})$ in a partial model $\mathfrak{M}$ just in case the two sentences have the same truth value (or lack thereof) in $\mathfrak{M}$. Let's say that a partial model $\mathfrak{M}$ has the *self-reference property* if for each formula $\varphi(x)$ there is a sentence γ that says $\varphi(\text{`}\gamma\text{'})$ both in $\mathfrak{M}$ and in all expansions of $\mathfrak{M}$. The proof of Gödel's self-reference lemma doesn't carry over to the present case, due to the excessive generality of our definition of 'valuation scheme'; so it is not sufficient to assume that the substitution relation is definable in $\mathfrak{M}$. Nonetheless, there is a rather simple way to assure that a partial model has the self-reference property.

LEMMA 3. *Let $\mathfrak{M}$ be a partial model for a language $\mathcal{L}$ such that for each formula $\varphi(x)$ of $\mathcal{L}_M$, there is a term t of $\mathcal{L}_M$ that denotes $\varphi(t)$ in $\mathfrak{M}$; then $\mathfrak{M}$ has the self-reference property.*

PROOF. This is almost, but not quite, as trivial as it looks. Let $\mathcal{L}$ be the language $\mathfrak{M}$ interprets, and let $\mathfrak{M}'$ be any expansion of $\mathfrak{M}$ to any $\mathcal{L}' \supseteq \mathcal{L}$ (where $\mathcal{L}'$ may simply be $\mathcal{L}$, in which case $\mathfrak{M}'$ is simply $\mathfrak{M}$). Given $\varphi(x)$, let $\gamma = \varphi(t)$, where t is a term

[8] In [**McG91**, Chapter 8]. See specifically his scheme σ_4.

that denotes $\varphi(t)$ in $\mathfrak{M}$, and hence in $\mathfrak{M}'$. Then the terms t and '$\varphi(t)$' denote the same thing, so by clause 3 of the definition of 'valuation scheme' the sentences $\varphi(t)$ and $\varphi(\text{`}\varphi(t)\text{'})$ are both true or both false in $\mathfrak{M}'$—in other words, $\gamma$ says $\varphi(\text{`}\gamma\text{'})$. □

From now on we will assume that $\mathfrak{M}$ has the self-reference property. Now let λ be any sentence that says $\neg T(\text{`}\lambda\text{'})$ (in all expansions of $\mathfrak{M}$); it is trivial to show that λ cannot have a truth value in any fixed point over $\mathfrak{M}$.

Sentences like λ—those that are not true or false in any fixed point over $\mathfrak{M}$—are called *paradoxical* in $\mathfrak{M}$. A weaker feature than paradoxicality is ungroundedness: a sentence is *ungrounded* in $\mathfrak{M}$ if it receives no truth value in the *least* fixed point over $\mathfrak{M}$. An example of a sentence that is ungrounded but not paradoxical is a *truth-teller*, a sentence τ that says $T(\text{`}\tau\text{'})$.

We shouldn't get too caught up in the euphoria of all these positive results, however: Kripke's approach, like Tarski's, has its limitations. Certain notions turn out to be indefinable in partial models in much the way truth is indefinable in classical models. In particular, while *true* and *false* are expressible in many partial models, the same cannot be said for *not true*. To see this, let $\mathfrak{M}$ be any partial model, and suppose $\nu(x)$ is a formula that defines $\{\varphi : \varphi \text{ is not true in } \mathfrak{M}\}$. Then if λ is a sentence that says $\nu(\text{`}\lambda\text{'})$, then $\lambda$ is true in $\mathfrak{M}$ iff $\nu(\text{`}\lambda\text{'})$ is true in $\mathfrak{M}$ (since λ says $\nu(\text{`}\lambda\text{'})$), iff λ is not true in $\mathfrak{M}$ (since $\nu(x)$ defines nontruth in $\mathfrak{M}$). This is a contradiction. Likewise, no partial model has its own unsatisfaction predicate: suppose to the contrary that $\upsilon(x, y)$ defined in $\mathfrak{M}$ the relation $U = \{(\varphi(x), a) : \varphi(\overline{a}) \text{ is not true in } \mathfrak{M}\}$, let $\psi(x)$ be the formula $\upsilon(x, x)$, and let γ be the sentence $\psi(\text{`}\psi(x)\text{'})$; then $\gamma$ is true iff $(\psi(x), \psi(x))$ belongs to the relation $U$, iff $\psi(\text{`}\psi(x)\text{'})$ is not true, i.e., iff γ is not true.

There are two points about this limitation that I want to bring out. One is that it is not simply a matter of *some notion or other* being inexpressible in a partial model. The inexpressible notions just described are ones we ourselves use in talking about partial models themselves—for example, we used the notion of untruth in a

model when we defined 'paradoxical' and 'ungrounded'. This poses a challenge, to say the least, to any view that claims that partial models model natural languages, or at least the whole of any natural language: how can they model the language we speak if, in describing these partial models in the first place, we use concepts that can't be expressed in them? It may be possible to meet this challenge; for now I am only concerned to point out that the challenge exists, and is potentially quite serious.

The second point is that this limitation is by no means specific to partial models: within a fairly broad set of constraints, *any* kind of formalized language will suffer a similar limitation. Namely, suppose we have specified some range of objects, call them *languages*; and for each language, we have defined another range of objects, call them *formulas*; and that we have defined, for each language $\mathcal{L}$ and each formula φ of $\mathcal{L}$ a class E, which we may call φ's *extension* in $\mathcal{L}$. (It should be clear that the terms 'language', 'formula' and 'extension' are merely placeholders.) Now suppose further that we have specified some particular language $\mathcal{L}$. Consider the class

$$H = \{\varphi : \varphi \text{ is a formula of } \mathcal{L} \text{ and } \varphi \text{ is not an element of } \varphi\text{'s extension in } \mathcal{L}\}$$

Obviously such an H exists, provided the formulas are eligible for class membership (generally a safe assumption). Now it is quite easy to show that H cannot be the extension of any formula of $\mathcal{L}$. What's more, $\mathcal{L}$ is by assumption a language we have specified, and H was defined in terms of $\mathcal{L}$ using the terms 'formula' and 'extension' which, by assumption, we understand; so there can be no question of our having specified H in *our* language. So some class specifiable in our language is not specifiable in $\mathcal{L}$, at least in the sense that it is not the extension of any formula of $\mathcal{L}$. The challenge that faces partial models' claims to correctly model natural language faces any formalized language's claim to do so.

2. Proof Theory

Most of the formal work on the Liar has focused on formal languages, and in particular on questions of definability. However, the Liar paradox is first and foremost a pattern of reasoning, which seems correct but has an unacceptable conclusion. It therefore seems reasonable to bring the methods of logic to bear directly on these reasoning patterns, if only to see precisely what assumptions are needed to generate a contradiction. Accordingly, in this section we will investigate several *formal deductive systems*, which serve as idealized models of the reasoning we engage in much as formalized languages serve as models of the languages we speak.

We will begin with a sound and complete system for classical propositional logic. Some of our ideas will be expressed most effectively if we choose a Gentzen-style system.[9] In such a system, a proof is not a collection of formulas but of *sequents*; a sequent is a finite set Γ of sentences, followed by a separator symbol '$\vdash$', followed by a single sentence φ. The idea is that the sentences in Γ are the assumptions under which φ is being asserted; a derivation of $\Gamma \vdash A$ is considered a proof of φ from the sentences in Γ. This mirrors the natural language practice of reasoning hypothetically, making assumptions to see what follows. We will make certain standard notational abbreviations when writing down sequents: braces ($\{$, $\}$) around the elements of Γ will be omitted, and we will write Γ, φ (or φ, Γ) for $\Gamma \cup \{\varphi\}$; similarly, we will write Γ, Δ for $\Gamma \cup \Delta$. We also write $\vdash \varphi$ for $\emptyset \vdash \varphi$.

Before actually presenting a particular system, let us make some general definitions. A *deductive system* is a set $\mathcal{S}$ of binary relations between finite sets of sequents and individual sequents; these relations are called the *rules* of $\mathcal{S}$. A set S of sequents is closed under a rule R if, whenever $(X, \Sigma) \in R$ and $X \subseteq S$, then $\Sigma \in S$; a sequent is *derivable* in $\mathcal{S}$ iff it belongs to the smallest set of sequents closed under the rules of $\mathcal{S}$. A sequent Σ is called an *initial sequent* of $\mathcal{S}$ if $(\emptyset, \Sigma) \in R$ for some rule R of $\mathcal{S}$.

[9]The present treatment of deductive logic is based loosely on that of [**Tak75**].

We specify a rule R as follows:

$$R: \quad \frac{\Sigma_1 \ldots \Sigma_n}{\Sigma}$$

indicates that $(\{\Sigma_1 \ldots \Sigma_n\}, \Sigma)$ belongs to R. We say that a *sentence* φ is *provable* in $\mathcal{S}$ just in case the sequent $\vdash \varphi$ is derivable in $\mathcal{S}$; φ is provable *from* $\psi_1 \ldots \psi_n$ in $\mathcal{S}$just in case the sequent $\psi_1 \ldots \psi_n \vdash \varphi$ is derivable in $\mathcal{S}$.

With these generalities out of the way we can consider some specific deductive systems. For simplicity I will only consider sentential logic, i.e., the following systems will not have any rules for quantifiers. Extending the following results to full predicate logic is straightforward, but requires some tedious detail that would only distract us from our purposes. (We *do* allow quantifiers to appear in the sequents of our system; we just don't bother to include enough rules to prove all the quantifier inferences.) Let us assume as given a fixed language $\mathcal{L}$ with a distinguished unary predicate T; it is convenient to take the connectives $\neg$, $\vee$ and $\supset$ as primitive. (However, '$\varphi \rightarrow \psi$' still abbreviates '$\neg\varphi \vee \psi$'.)

The rules of our first system are as follows:

identity: $\varphi \vdash \varphi$

weakening: $\dfrac{\Gamma \vdash \psi}{\varphi, \Gamma \vdash \psi}$ cut: $\dfrac{\Gamma \vdash \varphi \quad \varphi, \Delta \vdash \psi}{\Gamma, \Delta \vdash \psi}$

$\supset$ left: $\dfrac{\Gamma \vdash \varphi \quad \psi, \Delta \vdash \chi}{\varphi \supset \psi, \Gamma, \Delta \vdash \chi}$ $\supset$ right: $\dfrac{\varphi, \Gamma \vdash \psi}{\Gamma \vdash \varphi \supset \psi}$

$\vee$ left: $\dfrac{\varphi, \Gamma \vdash \chi \quad \psi, \Delta \vdash \chi}{\varphi \vee \psi, \Gamma, \Delta \vdash \chi}$ $\vee$ right: $\dfrac{\Gamma \vdash \varphi}{\Gamma \vdash \varphi \vee \psi}$ and $\dfrac{\Gamma \vdash \psi}{\Gamma \vdash \varphi \vee \psi}$

$\neg\neg$ left: $\dfrac{\varphi, \Gamma \vdash \psi}{\neg\neg\varphi, \Gamma \vdash \psi}$ $\neg\neg$ right: $\dfrac{\Gamma \vdash \varphi}{\Gamma \vdash \neg\neg\varphi}$

$$\neg\supset \text{ left: } \begin{cases} \dfrac{\varphi, \Gamma \vdash \chi}{\neg(\varphi \supset \psi), \Gamma \vdash \chi} \\ \\ \dfrac{\neg\psi, \Gamma \vdash \chi}{\neg(\varphi \supset \psi), \Gamma \vdash \chi} \end{cases} \qquad \neg\supset \text{ right: } \frac{\Gamma \vdash \varphi \quad \Delta \vdash \neg\psi}{\Gamma, \Delta \vdash \neg(\varphi \supset \psi)}$$

$$\neg\vee \text{ left: } \begin{cases} \dfrac{\neg\varphi, \Gamma \vdash \chi}{\neg(\varphi \vee \psi), \Gamma \vdash \chi} \\ \\ \dfrac{\neg\psi, \Gamma \vdash \chi}{\neg(\varphi \vee \psi), \Gamma \vdash \chi} \end{cases} \qquad \neg\vee \text{ right: } \frac{\Gamma \vdash \neg\varphi \quad \Delta \vdash \neg\psi}{\Gamma, \Delta \vdash \neg(\varphi \vee \psi)}$$

ex absurdo quodlibet: $\varphi, \neg\varphi \vdash \psi$ excluded middle: $\vdash \varphi \vee \neg\varphi$

Let us call this system $\mathcal{G}$ (after Gentzen). The identity, weakening and cut rules are called *structural rules*, and the others are called *connective rules*. $\mathcal{G}$ is sound and complete in the sense that a sequent $\Gamma \vdash \varphi$ is derivable in it just in case φ is a classical tautological consequence of Γ.

Given a subset S of $\{\neg, \vee, \supset\}$, let $\mathcal{L}_S$ be the set of sentences of $\mathcal{L}$ not containing any connectives outside S. Call a system *sound and complete for* $\mathcal{L}_S$ if a sequent $\Gamma \vdash \varphi$ made of sentences of $\mathcal{L}_S$ is derivable just in case φ is a tautological consequence of Γ. If S does not contain $\neg$, we can obtain a sound and complete system for $\mathcal{L}_S$ by taking as our rules the structural rules and, for each connective c in S, the rules c left and c right. This system can in turn be made into a sound and complete system for $\mathcal{L}_{S\cup\{\neg\}}$ by adding the $\neg\neg$ rules, *ex absurdo quodlibet* and excluded middle, and the $\neg c$ rules for each $c \in S$. Let us call the sound and complete system thus defined for S the system $\mathcal{G}_S$ (where S may or may not contain $\neg$). Thus $\mathcal{G}$ is simply $\mathcal{G}_{\neg,\vee,\supset}$.

Given a valuation scheme σ, we say that a set Γ of sentences *implies* a sentence A (relative to σ) just in case $\mathfrak{M} \models_\sigma \varphi$ for every partial model $\mathfrak{M}$ with $\mathfrak{M} \models_\sigma \bigwedge\Gamma$, where $\bigwedge\Gamma$ is the conjunction of Γ's elements.[10] A system $\mathcal{S}$ is *sound and complete* for

[10]An alternative definition of implication, due to Blamey, also requires that $\mathfrak{M} \neq\!|_\sigma \bigwedge\Gamma$ whenever $\mathfrak{M} \neq\!|_\sigma \varphi$. See [**Bla86**] for details.

σ just in case a sequent $\Gamma \vdash \varphi$ is derivable in $\mathcal{S}$ iff Γ implies φ relative to σ. Since Γ implies φ relative to the scheme vF just in case Γ implies φ classically, $\mathcal{G}$ is also sound and complete for vF. To find a system that is sound and complete for the scheme SK, it is convenient to consider not the full language $\mathcal{L}$ but the restricted language $\mathcal{L}_{\neg,\vee}$. Then an appropriate system is obtained by dropping excluded middle from $\mathcal{G}_{\neg,\vee}$; call the resulting system $\mathcal{G}_{\mathrm{SK}}$. Equivalently, $\mathcal{G}_{\mathrm{SK}}$ is obtained from $\mathcal{G}$ by dropping the $\supset$ and $\neg\supset$ rules and excluded middle.

There are some other interesting systems contained in $\mathcal{G}$ besides $\mathcal{G}_{\mathrm{SK}}$. For example, $\mathcal{G}_{\vee,\supset}$ is a fragment of intuitionistic logic; indeed, it is the complete intuitionistic logic of the language $\mathcal{L}_{\vee,\supset}$. (In intuitionistic logic, all of $\neg$, $\wedge$, $\vee$, and $\supset$ are primitive, as are both $\exists$ and $\forall$ in intuitionistic predicate logic.) Another interesting system is $\mathcal{G}_{\supset}$, or basic implicational logic.

In addition to its basic rules, a system has many derived rules. If the set of derivable sequents of $\mathcal{S}$ is closed under a rule R, we say that R is a *derived rule of* $\mathcal{S}$ *in the weak sense*. If the set of derivable sequents of $\mathcal{S}'$ is closed under R for any $\mathcal{S}'$ containing $\mathcal{S}$, we say that R is a derived rule of $\mathcal{S}$ *in the strong sense*.[11] For example, a famous result of Gentzen is that cut is a derived rule in the weak sense of the system $\mathcal{G}-$ cut; however, it is not a derived rule in the strong sense. Let us say that two systems $\mathcal{S}_1$ and $\mathcal{S}_2$ are *reformulations* of each other if every rule of one is a derived rule in the strong sense of the other. Thus reformulations are the same system for all intents and purposes. One reformulation of $\mathcal{G}_{\supset}$ is the system consisting of the rules identity, weakening, $\supset$ right, and

$$\textit{modus ponens}: \quad \frac{\Gamma \vdash \varphi \quad \Delta \vdash \varphi \supset \psi}{\Gamma, \Delta \vdash \psi}$$

For example, to see that any extension of $\mathcal{G}_{\supset}$ is closed under *modus ponens*, observe that a derivation of $\Gamma, \Delta \vdash \psi$ can be obtained from derivations of $\Gamma \vdash \varphi$ and $\Delta \vdash \varphi \supset$

[11]Equivalently, R is a derived rule of $\mathcal{S}$ in the strong sense if, whenever a sequent Σ follows from sequents $\Sigma_1 \dots \Sigma_n$ via R, then Σ is derivable in the result $\mathcal{S} + \{\Sigma_1 \dots \Sigma_n\}$ of adding $\Sigma_1 \dots \Sigma_n$ to $\mathcal{S}$ as initial sequents.

ψ, as follows:

$$\dfrac{\dfrac{\vdots}{\Delta \vdash \varphi \supset \psi} \qquad \dfrac{\dfrac{\vdots}{\Gamma \vdash \varphi} \qquad \dfrac{}{\psi \vdash \psi}\ \text{(identity)}}{\Gamma, \varphi \supset \psi \vdash \psi}\ (\supset \text{ left})}{\Gamma, \Delta \vdash \psi}\ \text{(cut)}$$

The rest of the proof that these two systems are reformulations of each other is similar, and is left to the reader.

So much for pure logic. To investigate proof-theoretic aspects of the Liar paradox, we need to add the means for self-reference and appropriate axioms for a truth predicate. The easiest way to achieve the former is as follows. Assume that the language $\mathcal{L}$ is countable and has an infinite number of individual constants, and let C_1 and C_2 be disjoint infinite sets of such. Let q be a 1–1 function from the sentences of $\mathcal{L}$ onto C_1, and write 'φ' instead of $q(\varphi)$. Let r be a 1–1 function from the formulas of $\mathcal{L}$ with one free variable onto C_2, and write $r(\varphi)$ as c_φ. The constant c_φ should be thought of as a name for the formula $\varphi(c_\varphi)$. Now let SR be the following system:

$$\text{ID}_1\text{:} \quad \vdash t = t \qquad \text{ID}_2\text{:} \quad s = t, \varphi(s) \vdash \varphi(t)$$

$$\text{Self-reference:} \quad \vdash c_\varphi = \text{`}\varphi(c_\varphi)\text{'}.$$

(In the rule ID_2, $\varphi(t)$ is just like $\varphi(s)$ except that some, all, or none of the occurrences of s in the latter are replaced by t.) The identity rules are included here, rather than in $\mathcal{G}$, simply for convenience.

Finally, the rules for truth:

$$\text{RT}_1\text{:} \ \frac{\vdash \varphi}{\vdash T(\text{`}\varphi\text{'})} \qquad \text{RT}_2\text{:} \ \frac{\vdash T(\text{`}\varphi\text{'})}{\vdash \varphi}$$

$$\text{RT}_3\text{:} \ \frac{\vdash \neg\varphi}{\vdash \neg T(\text{`}\varphi\text{'})} \qquad \text{RT}_4\text{:} \ \frac{\vdash \neg T(\text{`}\varphi\text{'})}{\vdash \neg\varphi}$$

It's easy to show for the three valuation schemes we considered in the last section that T defines truth in a partial model just in case its extension is closed under RT_1 and RT_2, and T is a truth predicate for that model just in case its extension is also

closed under RT_3 and RT_4.[12] Notice that there are no sentences to the right of any of the turnstyles. Let RT_1^*–RT_4^* be the result of allowing such sentences; for example, RT_1^* is the rule

$$\frac{\Gamma \vdash \varphi}{\Gamma \vdash T(\text{`}\varphi\text{'})}\,.$$

The difference between the starred and unstarred rules can be thought of as follows. The unstarred rules allow one to make the expected inferences about truth only when one is prepared to assert outright the sentence to which the rule is being applied, whereas the starred rules allow one to apply those inferences to sentences which one is only willing to assert on the basis of some hypothesis. Thus, the starred rules are closer to our actual use of the truth predicate.

We are almost ready to prove some consistency results. A system $\mathcal{S}$ is *weakly inconsistent* if there are sentences φ and $\neg\varphi$ that are both theorems of $\mathcal{S}$, and *strongly inconsistent* if every sentence is a theorem of $\mathcal{S}$. (Obviously any strongly inconsistent system is weakly inconsistent, and the converse holds for any system containing *ex absurdo quodlibet* and cut, but does not hold generally.) Likewise, a system is *strongly* (*weakly*) *consistent* if it is not weakly (strongly) inconsistent.

THEOREM 5. *The system* $\mathcal{S} = \mathcal{G} + \mathrm{SR} + \mathrm{RT}_1$–$\mathrm{RT}_4$ *is* (*strongly*) *consistent.*

PROOF. Let $\mathfrak{M}_0$ be a classical or partial model for $\mathcal{L}$ that interprets each constant $\text{`}\varphi\text{'} = q(\varphi)$ as the sentence φ itself, and each constant c_φ as the sentence $\varphi(c_\varphi)$. Let $\mathfrak{M} = (\mathfrak{M}_0, (E, A))$ be any supervaluational fixed point over $\mathfrak{M}_0$. Let S be the set of sequents $\Gamma \vdash \varphi$ such that the conditional $\bigwedge \Gamma \supset \varphi$ is true in $\mathfrak{M}$. (If Γ is empty, then $\bigwedge \Gamma \supset \varphi$ is simply the sentence φ.) A routine verification shows that S is closed under the rules of $\mathcal{G}$, and S is closed under SR by our choice of $\mathfrak{M}_0$. Finally, since $\vdash \varphi$ belongs to S just in case φ is true in $\mathfrak{M}$, S is closed under RT_1–RT_4 by the definition of 'fixed point'. Thus S is closed under the rules of $\mathcal{S}$, and therefore contains every

[12]Or more accurately, just in case the set $\{\vdash \varphi : \varphi$ belongs to T's extension$\}$ is closed under these rules.

sequent derivable in $\mathcal{S}$. If $\vdash \varphi$ and $\vdash \neg\varphi$ were derivable in $\mathcal{S}$, then φ and $\neg\varphi$ would both be true in $\mathfrak{M}$, which is impossible. □

Thus the weak rules RT_1-RT_4 are safe. Not so the starred ones:

THEOREM 6. *Let $\mathcal{S}$ be any system that extends $\mathcal{G}_\supset + \mathrm{SR} + \mathrm{RT}_1 + \mathrm{RT}_2^*$; then $\mathcal{S}$ is strongly inconsistent.*

PROOF. Let φ be any sentence, let $\psi(x)$ be the formula $T(x) \supset \varphi$, and let $c = c_\psi$. Then the sequent $\vdash A$ may be derived as follows:

1	$T(c) \vdash T(c)$	(identity)
2	$\vdash c = \text{`}T(c) \supset \varphi\text{'}$	(self-reference)
3	$T(c) \vdash T(\text{`}T(c) \supset \varphi\text{'})$	(from 1 and 2)
4	$T(c) \vdash T(c) \supset \varphi$	(RT_2^*)
5	$T(c) \vdash \varphi$	(from 1 and 4 by *modus ponens*)
6	$\vdash T(c) \supset \varphi$	($\supset$ right)
7	$\vdash T(\text{`}T(c) \supset \varphi\text{'})$	(RT_1)
8	$\vdash T(c)$	(from 2 and 7)
9	$\vdash \varphi$	(from 6 and 8 by *modus ponens*)

□

This version of the Liar is called *Curry's paradox*. (Versions of Theorem 6 using the more familiar version of the Liar also exist.) What's striking about Curry's paradox is that the underlying logic needed to generate the contradiction is very minimal: essentially, RT_1 and RT_2^* yield a contradiction in the presence of any connective obeying *modus ponens* and $\supset$ right. In particular, the paradox is not tied in any essential way to classical proof theory. (The system $\mathcal{G}_\supset$ is a subsystem of intuitionistic logic, for example.) Another notable feature of Curry's paradox is the fact that the inconsistency obtained is *strong* inconsistency, despite the fact that not every weakly inconsistent system extending $\mathcal{G}_\supset$ is strongly inconsistent; so the fact that the system

$\mathcal{S}$ has every sentence as a theorem is not due to the classical principle that everything follows from a contradiction.

To push this last point further, notice that Theorem 6 remains true when $\mathcal{G}_{\supset}$ is replaced by $\mathcal{G}_{\supset}^{-}$, the result of restricting the modus ponens-based formulation of $\mathcal{G}_{\supset}$ to sequents with at most one formula to the left of the turnstyle. $\mathcal{G}_{\supset}^{-}$ is contained in relevance logic;[13] so even a relevantist who accepted RT_1 and RT_2^* would be committed to every sentence, however "irrelevant."

An analogue of Theorem 6 applies to the system $\mathcal{G}_{\neg,\vee}$ (the proof is straightforward and is left to the reader):

THEOREM 7. *Any system extending* $\mathcal{G}_{\neg,\vee}$+SR+RT_1^*+RT_2^* *is strongly inconsistent.*

□

Neither of these results applies to $\mathcal{G}_{\mathrm{SK}}$, as the following shows:

THEOREM 8. *The system* $\mathcal{S} = \mathcal{G}_{\mathrm{SK}} + \mathrm{SR} + \mathrm{RT}_1^*$–$\mathrm{RT}_4^*$ *is consistent.*

PROOF. Similar to the proof of Theorem 5. Let $\mathfrak{M}_0$ be as in the proof of that theorem, and let $\mathfrak{M} = (\mathfrak{M}_0, (E, A))$ be a fixed point over $\mathfrak{M}_0$ relative to the strong Kleene scheme. Let S be the set of sequents $\Gamma \vdash \varphi$ such that if all the elements of Γ are true in $\mathfrak{M}$, then φ is also true in $\mathfrak{M}$. Inspection of the rules of $\mathcal{G}_{\mathrm{SK}}$ shows that S is closed under these rules, and S is closed under the rules of SR by our choice of $\mathfrak{M}_0$. Closure under RT_1^*–RT_4^* follows from the fact that $\mathfrak{M}$ is a fixed point. For example, to see that S is closed under RT_1^*, suppose $\Gamma \vdash \varphi \in S$. If not every sentence of Γ is true in $\mathfrak{M}$, then $\Gamma \vdash T(\text{`}\varphi\text{'}) \in S$ trivially, so suppose every sentence of $\Gamma$ is true in $\mathfrak{M}$. Then $\varphi$ is true in $\mathfrak{M}$, so (since $\mathfrak{M}$ is a fixed point) $T(\text{`}\varphi\text{'})$ is true in $\mathfrak{M}$ also, and therefore $\Gamma \vdash T(\text{`}\varphi\text{'}) \in S$. Closure under RT_2^*–RT_4^* is proven similarly. □

[13] At least in some of its formulations; I am thinking here of the system **E** of [**AB75**]. It should be noted that their Gentzen-style formulation of **E** has neither *modus ponens* nor cut as a basic rule (though both are derived rules in the weak sense); however, *modus ponens is* a basic rule in both their axiomatic and their natural deduction formulation of **E**. So including it in a Gentzen-style formulation of relevance logic seems reasonable.

REMARK. It may not be obvious why we can't use this proof to show that $\mathcal{G}+$ SR + RT_1^*–RT_4^* is consistent, by letting $\mathfrak{M}$ be a supervaluational fixed point rather than a strong Kleene fixed point. The reason is that although S would then have been closed under excluded middle, it would not have been closed under $\vee$ left. In particular, it is possible to have φ and ψ undefined, $\varphi \vee \psi$ true (since a true disjunction need not have any true disjuncts in vF), and χ false, in which case $\varphi \vdash \chi$ and $\psi \vdash \chi$ will belong to S but $\varphi \vee \psi \vdash \chi$ will not.

In contrast to the preceding positive result, we have the following negative result. Here c is c_φ, where $\varphi(x)$ is the formula $\neg T(x)$.

THEOREM 9. *A system $\mathcal{S}$ is weakly inconsistent if it contains the system* SR*, and if either (a) $T(c)$ is a theorem of $\mathcal{S}$ and $\mathcal{S}$ contains* RT_2 *or (b) $\neg T(c)$ is a theorem of $\mathcal{S}$ and $\mathcal{S}$ contains* RT_1*.*

PROOF. Assuming (a) holds, we have the following derivation in $\mathcal{S}$:

1	$\vdash T(c)$	
2	$\vdash c = \text{`}\neg T(c)\text{'}$	(self-reference)
3	$\vdash T(\text{`}\neg T(c)\text{'})$	(1 and 2, ID_2)
4	$\vdash \neg T(c)$	(3, RT_2)

and so both $T(c)$ and $\neg T(c)$ are theorems of $\mathcal{S}$, so $\mathcal{S}$ is weakly inconsistent. The proof is similar if (b) holds. □

Thus, adding an axiom that classifies the Liar sentence $T(c)$ as true or untrue tends to make a consistent system inconsistent.

CHAPTER 3

Truth Value Gap Accounts

In Chapter 1, we encountered the view that Liar sentences are neither true nor false and, hence, not true. We also saw there, as well as in the last chapter, that this view is of limited value in explaining or "solving" the Liar. In particular, we saw that in the presence of (T), or even a certain restriction of (T), the claim that the Liar is not true leads directly to a contradiction, almost regardless of what the underlying logic is.[1]

Similarly, it is possible to derive from the claim that the Liar is not true a contradictory-looking (but not formally inconsistent) conclusion, using almost no logical assumptions and *no* assumptions about truth whatsoever. Namely, begin by considering the sentence

(1) (1) is not true

and assume, with the gap account, that

(2) (1) is not true.

Using the fact that (1) = '(1) is not true', this gives us

(3) '(1) is not true' is not true.

Finally, using (2) and (3), we have

(4) (1) is not true and '(1) is not true' is not true.

[1]Strictly speaking, of course, to derive a contradiction we also need to use an identity such as $S =$ 'S is not true'. I usually won't bother to make this sort of qualification explicitly.

But (4) is of the form 'A, and 'A' is not true', which just looks like a contradiction. While it is not formally a contradiction, absent any principles about truth, it doesn't look like the sort of thing one should go around asserting.

The gap account therefore faces some serious challenges. Interestingly, there are views based on the gap approach that at least appear to avoid the abovementioned problems. Since asserting anything about whether the Liar is true seems to get us into trouble rather quickly, these views all, in one way or another, provide principled reasons for not making such assertions.

1. Truth and Determinate Truth

This modified gap view maintains that the predicate 'true' is *partially defined*, in the sense that there are some sentences such that it would be incorrect to say that they are true, and equally incorrect to say that they are *not* true. In this respect, 'true' is similar to *vague* predicates, like 'poor' and 'bald': suppose Larry is on the borderline between bald and nonbald. Then it would certainly be wrong to say that Larry is bald; but it would be equally wrong to say that Larry is not bald. Both claims, that Larry is bald and that Larry is not bald, are mistaken and must be rejected. Likewise, on this view, both the claim that (1) is true and the claim that (1) is not true must be rejected. (And "rejecting" a statement is not the same as asserting its negation, since if it were, we would be claiming that Larry is not bald when we reject the claim that Larry is bald, which is precisely what we mustn't do.) According to some proponents of this view (e.g., McGee), this is tantamount to saying that 'true' is *vague*; in that case, sentences like (1) are *borderline cases* of truth in much the way Larry is a borderline case of baldness. Others (e.g., Soames) refrain from calling 'true' vague, preferring instead to say that like vague predicates, it is *partially defined*.[2]

[2]The presentation that follows is based loosely on [**Soa97**].

This view certainly avoids the strengthened Liar problem, at least in the form it takes in connection with the simple gap account. The problem there was that from an assertion that (1) is not true we can argue that (1) is true after all; the argument just doesn't get going if, instead of claiming that (1) is not true, we refuse to assert either that it is true or that it isn't. But the vagueness account would be *ad hoc* if it offered no reason (or no other reason) to think that 'true' is partial.

Fortunately, it does provide such a reason. It does this by providing a picture, which one might believe independently of the Liar paradox, of how 'true' gets its meaning. Our linguistic competence in using 'true' (the story goes) consists in our having accepted some sort of disquotational rule or set of rules. Those rules might be RT_1–RT_4 of the last chapter, for example. The conventions of language put a sentence A into the extension of 'true' just in case 'A is true' can be derived from the rules governing 'true' (in some suitable sense of 'derived') together with whatever other sentences the conventions of language (together with the world) settle as true. This will include 'snow is white', ''snow is white' is true', etc., but it will *not* include ungrounded sentences like (1). The result is a truth predicate resembling the least fixed point in Kripke's construction, as described in the last chapter.

This is a compelling picture, at least at first blush. It also illustrates what's wrong with a certain kind of objection to some accounts of the paradoxes. In his [**Chi79**], Charles Chihara criticizes theories of truth such as Kripke's when they are understood as descriptive accounts of our ordinary concept of truth, on the grounds that those theories are too complicated to be theories of what we ordinarily mean by 'true'. Or as Chihara puts it, "[these theories have] suggested only very complex and logically sophisticated notions of 'true'–notions which no one (certainly not ordinary speakers of English) could be expected to have learned as children." (p. 610) The trouble with this is that it construes Kripke's theory of truth and others like it as *analyses* of the notion of truth, rather than as *descriptions* of (a fragment of) a language. The present case illustrates the difference nicely. A really worked-out

version of the partiality account, in which the notion of something's being "settled" by the conventions of language is made precise, might use Kripke's mathematical machinery, or perhaps some other, more complicated machinery. But in employing such machinery we are not attributing mastery of it to ordinary speakers of English. Rather, all it takes, on this account, to be a competent user of the word 'true' is mastery of (say) the rules RT_1–RT_4.

Unfortunately, the partiality account has serious problems. Some of these come to light when we look closely at precisely what disquotational rules we are supposed to have adopted. I assumed for the sake of example that they were RT_1–RT_4; this was not entirely arbitrary, since we saw in the last chapter that if the underlying logic is classical, the stronger rules RT_1^*–RT_4^* are inconsistent. If the choice of RT_1–RT_4 is taken seriously, then the partiality account is claiming that while an inference from 'S' to ''S' is true' or vice versa is licensed as part of the meaning of 'true' when the premise of the inference has been asserted unconditionally, it is not so licensed otherwise.

This seems wrong. In the course of ordinary reasoning, we regard either inference as valid whether or not we are prepared to assert 'S' unconditionally, and likewise for the other rules. For example, the argument

> Everything the meteorologist says is true; therefore, if he says that it will snow tomorrow, then it will snow tomorrow

is intuitively valid, yet the conclusion cannot be derived from the premises using just the (appropriate informal counterparts of the) rules RT_1–RT_4. There is just no clear sense in which we have accepted RT_1–RT_4 but not RT_1^*–RT_4^*.

Now this is not the end of the story, since where truth value gaps are involved, the inference rules of classical logic seem negotiable. Some philosophers favor treating vague predicates supervaluationally, which makes the underlying proof theory classical; but a more popular approach is to treat them by means of the strong Kleene

valuation scheme, which yields a different proof theory. And as we saw, RT_1^*–RT_4^* are perfectly consistent with that proof theory.

What logic best characterizes our reasoning in the presence of vague predicates is not an easy question. On the one hand, people have a reluctance to endorse excluded middle when the constituent sentence is gappy: if a color sample is intermediate between blue and turquoise, for example, many people will strongly resist the claim that it's either blue or not blue. On the other hand, those same people will very likely agree that it's not both blue and not blue, despite the fact that $\varphi \vee \neg\varphi$ and $\neg(\varphi \wedge \neg\varphi)$ are logically equivalent both classically *and* in the strong Kleene scheme. Moreover, it has often been complained[3] that nothing like sustained ordinary reasoning is possible in strong Kleene logic.

In any case, there is strong reason to think that at the very least some extension of strong Kleene logic is required to adequately model ordinary reasoning. The main reason is that we often use natural language to express so-called *penumbral connections* between vague predicates. As an example, canary is evidently a shade of yellow, a fact we can partially express by saying

(5) Every canary object is a yellow object.

Yet both 'canary' and 'yellow' are vague, and something might be a borderline case of both predicates (perhaps because it is intermediate between yellow and brown). For this reason, (5) cannot be rendered formally as $\forall x\,(C(x) \to Y(x))$ if '$\forall$' and '$\to$' have their usual interpretations and the strong Kleene scheme is adopted: that would require $C(x) \to Y(x)$ to be true of all objects, including our borderline canary object.

One way to accommodate penumbral connections would be to introduce a new conditional '$\supset$', with a different interpretation from '$\to$', and express (5) as $\forall x\,(C(x) \supset Y(x))$. Doing so would very likely render RT_1^*–RT_4^* inconsistent. One way to see this is that the truth definition of the resulting language would certainly be nonmonotone, since

[3] e.g., in [**Fef82**]

(as we have seen) a sentence $\varphi \supset \psi$ can be true when its antecedent and consequent are gappy, but obviously must be false when its antecedent is made true and its consequent false. And since Kripke's fixed-point results all require a monotone truth definition, those results are unlikely to apply to a language containing '$\supset$', and indeed are known not to apply when '$\supset$' is given a reasonable truth table definition.

A more proof theoretic way of making this point is to observe that '$\supset$' is likely to satisfy *modus ponens* and the rule of *conditional proof*: if B has been proven from the assumption A, then 'if A then B' may be inferred without the assumption A. More importantly for my purposes, it is highly plausible that English has *some* conditional that obeys these rules. Now the rule of conditional proof is really nothing other than the rule $\supset$ right; and in the last chapter we saw that RT_1^*–RT_4^* are jointly inconsistent in the presence of any connective obeying these rules.

The other main problem with the partiality account is that it doesn't really avoid the strengthened Liar problem. We have spoken thus far about sentences that are "settled" or left "unsettled" by the conventions of language, and of the settled ones as being settled as true or untrue. Let us call those sentences that are settled as true "determinately true," those settled as false "determinately false," and call those left unsettled "indeterminate." We are not introducing any new concepts here; we are only giving names to concepts we have been employing throughout this entire discussion. According to the partiality account, then, paradoxical sentences like (1) are indeterminate, and therefore not determinately true. Now consider the following "determinate Liar" sentence:

(D) (D) is not determinately true.

A proponent of the vagueness account would presumably say that (D) is indeterminate–in any event, he would certainly deny that it is determinately true. But to do so is to assert that (D) is *not* determinately true. And if (D) is not determinately true, then it surely follows that '(D) is not determinately true' is determinately true.

But from *this* it follows immediately that (D) is determinately true, contradicting the assumption we began with. Even if we reject the inference from '(D) is not determinately true' to ''(D) is not determinately true' is determinately true, if we assert that (D) is not determinately true then we are committed, without any assumptions about determinate truth whatsoever, to both

(D) is not determinately true

and

'(D) is not determinately true' is not determinately true

i.e., we are committed to a statement of the form 'S, and 'S' is not determinately true'. And this seems just as contradictory as the analogous statement about truth from the start of this chapter.

The situation is just like that described earlier, with 'true' replaced by 'determinately true' and 'false' replaced by 'determinately untrue'. When T is interpreted as 'determinately true', RT_1–RT_4 still seem valid, and will lead to a contradiction in the presence of either '(D) is determinately true' or '(D) is not determinately true'.

It may be worth noting that the rules intuitively governing determinate truth that we've appealed to do not themselves generate a formal contradiction; this is seen from the fact, proved in the last chapter, that RT_1–RT_4 are classically consistent. They do, however, generate a certain instability. We have just seen that we must reject both the claim that (D) is determinately true and the claim that (D) is not determinately true, on pain of contradiction. One would have thought that one of these two options must obtain. But in any case, we would also contradict ourselves by saying that '(D) is determinately true' is indeterminate: for that would imply that '(D) is not determinately true' is also indeterminate, which is just to say that (D) is indeterminate and, hence, not determinately true; but as we know, we must not say that (D) is not determinately true, on pain of contradiction. So even if the

rules intuitively governing determinate truth do not by themselves generate a formal contradiction, they do generate what looks like an informal incoherence.

Moreover, a formal contradiction will arise if we assume the plausible schema

(6) If 'A' is determinately true, then A

(The 'if' here is the usual material conditional; all that is essential to the argument, however, is that it supports the inference from 'if A then not A' to 'not A'.) Then we can argue as follows:

(1) (D) = '(D) is not determinately true'
(2) If '(D) is not determinately true' is not determinately true, then (D) is not determinately true (from (6))
(3) If (D) is determinately true, then (D) is not determinately true (1 & 2)
(4) (D) is not determinately true (3)
(5) '(D) is not determinately true' is determinately true (4)
(6) (D) is determinately true (1 & 5)

As for (6), it may not be as intuitively obvious as the determinate truth analogues of RT_1 and RT_2, though it is plausible and holds in partial models when a 'determinately' operator has been introduced in a standard way, at least on either the strong Kleene or the supervaluational approach.

In any event, the fact remains that the rules RT_1 and RT_2 are intuitively just as plausible when T is interpreted as 'determinately true' as they are when it is interpreted as 'true', and they render the partiality account of the Liar, as here stated, inconsistent. One way of handling this, due to Soames, will be discussed in the next chapter; another, due to McGee, we will examine next.

2. McGee's Response

Vann McGee criticizes the argument just given in his [**McG91**]. In particular, he rejects the move from '(D) is not determinately true' to ''(D) is not determinately

true' is determinately true'. More generally, he rejects *any* inference from

$$S \text{ is not determinately true}$$

to

$$\text{`}S\text{ is not determinately true' is determinately true.}$$

He writes:

> This argument is no good. From the hypothesis that a sentence is unsettled, it by no means follows that it has been settled that the sentence is unsettled. Quite the contrary, if a sentence is unsettled, then we are free to adopt linguistic conventions that settle it. (p. 7)

Although this thought is the intuitive basis for the technical work he goes on to do, he never elaborates on the remark itself. This is unfortunate, for it is really at the heart of the matter. In any event, I want to argue that this response won't work, that the determinate Liar sentence should be just as worrisome to someone who believes the truth predicate is vague as the ordinary strengthened Liar sentence is to someone who thinks the truth predicate is neither true nor false.

Before getting into the details of this, a certain objection must be dealt with. McGee's project is, in some sense, revisionary rather than descriptive. It is certainly revisionary in the sense that he is proposing a consistent replacement for our inconsistent naive theory of truth. But it is also in some sense a replacement for our ordinary *concept* of truth as well. McGee writes:

> I want to treat 'true' as a vague predicate. I do not intend to suggest by this that, in ordinary usage, 'true' is simply a vague predicate like ordinary vague predicates. Ordinary vague predicates are predicates whose applicability is underdetermined by the rules of our language, whereas, intuitively, our linguistic rules overdetermine the applicability of the word 'true' in conflicting ways. (pp. 7–8)

It may be objected that even if the arguments of the last section work against a descriptive account of 'true' as a vague predicate, they are inapplicable to McGee's account. Since McGee is doing conceptual revision, he is free to replace our ordinary concepts of truth (and, if necessary, of determinate truth) with cleaned-up concepts to which the intuitive principles adduced above do not apply. In particular, he is free to introduce a notion of determinate truth for which the inference from

$$S$$

to

$$\text{`}S\text{' is determinately true}$$

is not valid.

I have two things to say to this. First, if I can show that nothing in McGee can be used to avoid the arguments of the last section, then I will be content. But second, this defense of McGee is too simple. Although McGee is trying to effect a conceptual revision, not just any revision will do. If any revision would do, then it would be open to McGee to adopt a simple gap approach and reject the rules RT_1-RT_4. As we have seen, this would commit him to assertions of the form 'P, and 'P' is not true'. Given the ordinary meaning of 'true', these assertions seem contradictory and any theory that yields them seems self-defeating. But if we are allowed to give 'true' any meaning we like, then we can certainly give it a meaning which is such that there is no contradiction at all in such assertions.

McGee places more constraints on his project than that. He proposes the following criterion of success for any theory, and in particular for any theory of truth:

> (P2) A satisfactory theory should never make claims that are, according to the theory itself, untrue. (p. 5)

The project is not simply to assign a new concept to the word 'true': the new concept must genuinely be a concept of truth. McGee doesn't get much more precise than

this. Nonetheless, I think the project is clearly articulated enough that we can discuss what are and are not reasonable constraints on the theory of *determinate* truth that is presented, in addition to those on the theory of truth. I would maintain that insofar as (P2) is a reasonable criterion of success for a theory, the result of replacing 'untrue' in (P2) by 'not determinately true' is also a reasonable criterion of success. If this is right, then the arguments of the last section render any vagueness theory of truth inadequate.

Before examining McGee's remarks about determinate truth, we should examine the technical project that embodies those remarks. The book's main objects of study are what McGee calls *partially interpreted languages*. A partially interpreted language is a pair $(\mathfrak{A}, \Gamma)$, where $\mathfrak{A}$ is a classical model and Γ is a set of sentences of a language extending the language of $\mathfrak{A}$. Intuitively, we are to think of $\mathfrak{A}$ as the fully interpreted part of the language $(\mathfrak{A}, \Gamma)$ and of Γ as a set of *meaning postulates* which partially interpret the rest of the language. A sentence is *determinately true* in $(\mathfrak{A}, \Gamma)$ just in case it is a consequence of Γ together with the truths of $\mathfrak{A}$, in an appropriate sense of 'consequence'.

More precisely, let $\mathcal{L}$ be the set of all non-logical symbols that either occur in Γ or are interpreted by $\mathfrak{A}$. We define a *proof in* $\mathfrak{A}$*-logic* from a set S of sentences of $\mathcal{L}_A$ to be a sequence $\langle \varphi_\xi : \xi < \alpha \rangle$, for some ordinal α, of sentences of $\mathcal{L}_A$, such that each φ_ξ either (1) is a first-order valid sentence of $\mathcal{L}_A$, (2) is an element of S, (3) follows by *modus ponens* from previous sentences in the sequence, or (4) is $\forall x\, \psi(x)$ for some ψ, and for all $a \in A$, $\psi(\overline{a}) = \varphi_\upsilon$ for some $\upsilon < \xi$. (The sentence $\forall x\, \psi(x)$ of (4) is said to be justified by the $\mathfrak{A}$*-rule*, a generalization of the ω-rule.) A sentence φ of $\mathcal{L}_A$ is said to be *determinately true in* $(\mathfrak{A}, \Gamma)$ (in symbols: $(\mathfrak{A}, \Gamma) \vdash \varphi$) iff φ occurs in a proof in $\mathfrak{A}$-logic from $\Gamma \cup \{\text{true sentences of } \mathfrak{A}\}$.[4]

[4]McGee also defines an alternative definite truth relation, denoted $\models$, where $(\mathfrak{A}, \Gamma) \models \varphi$ iff every expansion of $\mathfrak{A}$ to the language of Γ that makes Γ true also makes φ true. $\vdash$ and $\models$ coincide when $\mathfrak{A}$ is countable, but do not coincide in general, and $\vdash$ is mathematically more tractable.

Although McGee does not use this terminology, we can define a partially interpreted language to be *consistent* just in case we do not have $(\mathfrak{A}, \Gamma) \vdash \varphi \wedge \neg\varphi$ for any φ. McGee's goal is to construct consistent languages that contain their own semantic notions; in particular, he seeks to construct a consistent partially interpreted language that contains predicates 'Tr' and 'Det' meaning *true* and *determinately true*, respectively. To this end, he provides adequacy conditions for a partially interpreted language to have its own truth and determinate truth predicates, and proves that under suitable conditions a partially interpreted language $(\mathfrak{A}, \Gamma)$ can be extended to a partially interpreted language $(\mathfrak{A}, \Gamma \cup \Delta)$ that satisfies those conditions, with Δ recursive. The adequacy conditions are complicated, and it is open to interpretation whether a language that satisfies them can be said to have its own determinate truth predicate. I will give just one of the conditions here:

$$\text{(C1)} \qquad (\mathfrak{A}, \Gamma) \vdash Det(\text{`}\varphi\text{'}) \quad \text{iff} \quad (\mathfrak{A}, \Gamma) \vdash \varphi.$$

Or with fewer symbols: a sentence φ is determinately true iff $Det(\text{`}\varphi\text{'})$ is determinately true. The following is *not* one of the adequacy conditions:

$$\text{(C2)} \qquad (\mathfrak{A}, \Gamma) \vdash \neg Det(\text{`}\varphi\text{'}) \quad \text{iff} \quad (\mathfrak{A}, \Gamma) \nvdash \varphi.$$

If a partially interpreted language satisfies both (C1) and (C2) then it is inconsistent. McGee proves this using complexity considerations, but the proof boils down to the following. Suppose (C1) and (C2) both hold, and let δ be a sentence such that $(\mathfrak{A}, \Gamma) \vdash \delta \leftrightarrow \neg Det(\text{`}\delta\text{'})$. If $(\mathfrak{A}, \Gamma) \nvdash \delta$, then by (C2) $(\mathfrak{A}, \Gamma) \vdash \neg Det(\text{`}\delta\text{'})$, and since $(\mathfrak{A}, \Gamma) \vdash \delta \leftrightarrow \neg Det(\text{`}\delta\text{'})$, it follows that $(\mathfrak{A}, \Gamma) \vdash \delta$, contrary to our assumption. So $(\mathfrak{A}, \Gamma) \vdash \delta$. But then $(\mathfrak{A}, \Gamma) \vdash Det(\text{`}\delta\text{'})$ by (C1), hence $(\mathfrak{A}, \Gamma) \vdash \neg\delta$ and therefore $(\mathfrak{A}, \Gamma) \vdash \delta \wedge \neg\delta$. Therefore, (C1) and (C2) cannot both hold for consistent languages. In fact, something further holds for these languages. *No* sentence of the form $\neg Det(\text{`}\varphi\text{'}) \wedge \neg Det(\text{`}\neg\varphi\text{'})$ is *ever* determinately true in a consistent language, regardless of whether Tr or Det occur in φ; and if φ is unsettled in a partially interpreted language (i.e.,

neither φ nor $\neg\varphi$ is determinately true in that language), then $\neg Det(\text{`}\varphi\text{'})$ is also unsettled in that language.

The failure of (C2) accords well with McGee's rejection of the inference from 'S is unsettled' to ''S is unsettled' is determinately true': while the sentence δ can be shown to be unsettled in the sense that neither it nor its negation is determinately true in any (consistent) language satisfying (C1), the formalization of 'δ is unsettled' is not determinately true in any of these languages, either.

I don't think McGee's partially interpreted languages can accurately be said to possess their own determinate truth predicates. Moreover, I think that the account of vagueness that they embody is inadequate, even as applied to cases having nothing to do with the semantic paradoxes. Intuitively, the determinate truths are the sentences that it would be correct to assert. Imagine for the moment that there is a population of speakers of $(\mathfrak{A}, \Gamma)$; the determinate truths of $(\mathfrak{A}, \Gamma)$ are those sentences that those speakers would be correct in asserting. Let φ be any sentence that is unsettled in $(\mathfrak{A}, \Gamma)$. As we have seen, $\neg Det(\text{`}\varphi\text{'})$ is not determinately true in $(\mathfrak{A}, \Gamma)$, i.e., the rules of that language do not license the assertion of $\neg Det(\text{`}\varphi\text{'})$. However, when we, who speak English, want to describe this situation, we *are* correct in asserting that φ is not determinately true in $(\mathfrak{A}, \Gamma)$–the rules of our language *do* license that assertion. It therefore seems that the sentence $\neg Det(\text{`}\varphi\text{'})$ is not an accurate translation into $(\mathfrak{A}, \Gamma)$ of the English sentence 'φ is not determinately true in $(\mathfrak{A}, \Gamma)$'.

The speakers of $(\mathfrak{A}, \Gamma)$ are therefore unable to express even the simplest facts about vagueness, namely the fact that a given sentence is indeterminate: even if '*Det*' did mean *determinately true in* $(\mathfrak{A}, \Gamma)$, they would never be correct in asserting a sentence $\neg Det(\text{`}\varphi\text{'}) \wedge \neg Det(\text{`}\neg\varphi\text{'})$. So even if φ is a translation into $(\mathfrak{A}, \Gamma)$ of 'Larry is bald', the speakers of $(\mathfrak{A}, \Gamma)$ have no way of correctly asserting that φ is unsettled. Moreover, they have no way of correctly asserting that δ is unsettled.

Something has surely gone wrong here. And since the objectionable features of partially interpreted languages are a natural outgrowth of McGee's view that whenever a sentence is unsettled it is unsettled that it is unsettled, that view is probably the source of the trouble. Indeed, the foregoing arguments can be applied directly to that claim. When I asserted earlier that 'Larry is bald' is indeterminate, that assertion was correct, just as it would have been *in*correct for me to assert that Larry is bald. But to say that my assertion was correct is simply to say that (in my ideolect at that time) ''Larry is bald' is indeterminate' is determinately true. Moreover, this argument works with 'Larry is bald' replaced by any other indeterminate sentence. It also shows that, if one is prepared to assert a sentence S, then one is committed to S's determinate truth, so that anyone who is prepared to assert that (D) is not determinately true is committed to the determinate truth of '(D) is not determinately true'.

So here's the situation. McGee has argued that whenever a sentence S is unsettled, so is the sentence 'S is unsettled'; I have argued for the opposite conclusion. If I am right, then McGee's argument goes wrong somewhere; but where? We can begin to answer this by unpacking his argument a bit. Recall that argument: if a sentence S is unsettled, then it is not settled that S is unsettled, since we are free to adopt conventions that settle S. In other words, if our current conventions settle 'S is unsettled' as true, then this constrains us from adopting any future conventions that settle S. Let's draw a distinction here. There are some conventions we could adopt such that adopting them would constitute a change in the meanings of some words, and there are others whose adoption would only constitute an increase in precision. McGee's argument should then read as follows: (1) If S is unsettled, then there are conventions we could adopt which would settle S and are such that adopting them would only constitute an increase in the precision of our terminology. (2) If 'S is unsettled' is determinately true, then any conventions we adopt that settle S would constitute a change in meaning. This seems to presuppose two principles:

(*) If 'S is unsettled' is determinately true at time t, then S is unsettled at time t

and

(M) If a sentence S is determinately true at time t, and if S is no longer determinately true at a later time t', then the change in conventions of language between t and t' constitutes a change of meaning, and not just an increase in precision.

((M) is reflected in the monotonicity property that determinate truth in partially interpreted languages enjoys: if $(\mathfrak{A}, \Gamma) \vdash \varphi$, and if $\Gamma \subseteq \Delta$, then $(\mathfrak{A}, \Delta) \vdash \varphi$.) McGee's argument now runs as follows:

> If S is unsettled at time t, then we are free to adopt new conventions at some later time t' that settle S without changing the meaning of any of our vocabulary. But if 'S is unsettled' is determinately true at t, and if meanings have not changed between t and t', then by (M), 'S is unsettled' is determinately true at t'. But then by (*) S is unsettled at t'. So if 'S is unsettled' is determinately true at t, then we are not free to adopt non-meaning-changing conventions that settle S after all, and therefore S is not unsettled. Contrapositively, if S is unsettled, then 'S is unsettled' is not determinately true.

This argument partly acknowledges, and partly ignores, the fact that determinate truth is relative to a language and to a time. If we are consistent in writing 'S is determinately true in $\mathcal{L}$ at t' instead of 'S is determinately true', then the above argument no longer works. For suppose S is unsettled in $\mathcal{L}$ at t, and 'S is unsettled' is determinately true in $\mathcal{L}$ at t; the latter really means that 'S is unsettled in $\mathcal{L}$ at t' is determinately true in $\mathcal{L}$ at t. Let t' be a later time, and assume no meaning-changing conventions are adopted between t and t', but that S is settled in $\mathcal{L}$ at t'. Then by (M), 'S is unsettled in $\mathcal{L}$ at t' is determinately true in $\mathcal{L}$ at t', and by (*), 'S is settled

in $\mathcal{L}$ at t'' is determinately true in $\mathcal{L}$ at t'. But there is no contradiction here: the sentences

$$S \text{ is unsettled in } \mathcal{L} \text{ at } t$$

and

$$S \text{ is settled in } \mathcal{L} \text{ at } t'$$

which are determinately true in $\mathcal{L}$ at t' are perfectly compatible, and indeed both seem correct.

Another, less formal way to see what's going on here is to notice that, intuitively, 'S is unsettled', 'S is determinately true', etc. all say something about the conventions of a language at a given time; so if those conventions change, they may change their determinate truth value, even if the change in conventions does not constitute a change in meaning.

Maybe the best way to understand 'S is unsettled' is as an *indexical* statement: 'S is unsettled' means something like 'S is unsettled in this language now'. Clearly, (M) does not hold in general for indexical sentences. In particular, if S is of the form 'The linguistic conventions of English have property P', then S can go from being determinately true in English to being determinately false in English when new conventions are adopted by English speakers, even if those new conventions do not change the meanings of any words. And sentences 'S is unsettled' are (synonymous with sentences) of this form, since they assert that the conventions of English render the assertion of S correct.

I therefore conclude that McGee's argument fails, and that someone who asserts a sentence S is committed to S's determinate truth. But in that case anyone who asserts that (D) is not determinately true is committed to the determinate truth of '(D) is not determinately true', and hence to the determinate truth of (D). The strengthened Liar problem is therefore just as much a problem for the vagueness account as it is for the truth value gap account, McGee's work notwithstanding.

CHAPTER 4

The Hierarchy Approach

1. Introduction

Next we will examine a set of views that seems to represent the most popular approach to the Liar at the moment. This approach starts from the idea that a sentence that ascribes truth can say different things on different occasions. Now no one denies *that*: a sentence 'S is true' is ambiguous whenever S is, for example. On the views we are about to consider, however, there is an additional context dependence, one which simultaneously blocks the strengthened Liar reasoning and explains why that reasoning nonetheless seems correct. Specifically, on such accounts, strengthened Liar reasoning depends on ignoring the contextually determined element that is present in truth ascriptions, and it is easily seen to be illegitimate once this element is made explicit. Aside from being popular, this approach also has a fairly long history; an early instance is Russell's [**Rus08**], and it has also been strongly influenced by the work of Tarski. I will call it the *hidden parameter* approach or, for reasons that will become apparent, the *hierarchy* approach.

Recall the strengthened Liar argument: let (S) be the sentence '(S) is not true', and suppose we have concluded that

(1) (S) is not true,

perhaps on the grounds that (S) is neither true nor false. We infer from (1) that

(2) '(S) is not true' is true

and conclude that

(3) (S) is true,

apparently contradicting (1). Hierarchy accounts accept this reasoning but deny that there is a contradiction. Sentences of the form 'A is true' may express different propositions in different contexts, and in particular the proposition expressed by (1) is not the negation of that expressed by (3).

It is worth noting here that once we attend to the context dependence of 'A is true', we can no longer regard truth as applying to sentence *types* (at least when those types themselves contain occurrences of 'true'): we must instead view it as applying to sentence tokens, or utterances, or types-in-contexts, or propositions, or the like. In particular, (1)–(3) above should be rewritten to reflect this. For definiteness, let us do so in terms of utterances. We start with an utterance of the form

(S′) (S′) is not true.

We then conclude

(1′) (S′) is not true.

At this point there are two moves available to us. On the one hand, we might say that (S′) and (1′) express different propositions, perhaps on the grounds that (S′) is self-referential in a way that (1′) isn't. In that case the proper analogue of (3) is not '(S′) is true' but rather '(1′) is true': clearly this is all we may legitimately infer from (1′). And clearly there is no inconsistency in this.

On the other hand, we might say that (S′) and (1′) *do* express the same proposition. In that case, we may infer as we did before that (1′) is true, but clearly we may conclude from this that (S′) is also true. So, as with the gap account, we have

an apparent contradiction: namely, between (1′) and

(3′) (S′) is true.

But as I noted above, (1′) need not express the negation of the proposition (3′) expresses.

At this point, we might recall from the previous chapter that there is another version of the strengthened Liar problem in which we show, with no assumptions whatever about truth and with very minimal assumptions about the underlying logic, that from the assumption that (S) is not true we may derive a statement of the form 'A, and 'A' is not true'. We begin with (1) and

(4) (S) = '(S) is not true'

and derive, by substitution of equals for equals,

(5) '(S) is not true' is not true

and from (4) and (5) we conclude

(6) (S) is not true and '(S) is not true' is not true.

And (6) just seems contradictory. How does the hierarchy account deal with this?

One thing we might call into question is the inference from 'A' and 'B' to 'A and B'. Once context sensitivity is taken seriously, we see that this rule is not valid in every situation. If A and B are asserted in different contexts, there might be no *single* context in which one is committed to *both* A and B, so one need not be committed to their conjunction. (If I say 'There is no beer', meaning there's no beer in the fridge, and later say 'There is beer', meaning there's beer at the store, I do not thereby commit myself to 'There is beer and there is no beer'.) However, I don't think that this consideration applies here. If we can arrange things so that '(S) is not

true' expresses the same proposition throughout the argument, then the argument ought to be valid; and there seems to be no obstacle to our so arranging things. In particular, if the shift in context is caused by the inference from 'A' to ''A' is true', then that shift in context will not occur in the present argument, since it uses no such rule.

So it looks like the hierarchy approach must accept the argument (4)–(6) as valid (at least in some contexts) and explain why (6), despite appearances, is not a contradiction. Probably the best way to do this is to say that (6) seems like a contradiction because its second conjunct formally contradicts what can legitimately be inferred from its first conjunct, namely ''(S) is not true' is true': what we forget when (6) strikes us as contradictory is that when ''(S) is not true' is true' is inferred from (6), a change of context has taken place and the inferred sentence expresses a proposition quite consistent with (6)'s second conjunct.

This chapter critically examines several versions of the hidden parameter/hierarchy approach, and develops a style of objection to them. Although each version suffers from problems particular to it, there is also a general problem they all suffer from. Basically, although they avoid one version of the strengthened Liar problem, they are all subject to more sophisticated versions of the problem. This is best seen by way of example, so let's look at a particularly simple example, the Tarskian hierarchy of languages.

In the last chapter we showed that no classical model in which substitution is definable has its own truth predicate. But nothing prevents our enlarging such a model $\mathfrak{M}$ to a model $\mathfrak{M}'$ that has a truth predicate for $\mathfrak{M}$, i.e., a $\mathfrak{M}'$ in which the set of truths of $\mathfrak{M}$ is definable. For example, if $\mathfrak{M}'$'s domain contains all subsets of $\mathfrak{M}$'s domain, and satisfies certain other conditions, then $\mathfrak{M}'$ automatically has a truth predicate for $\mathfrak{M}$; the proof of this mimics the Tarskian definition of truth given in the last chapter. Alternatively, we may form $\mathfrak{M}'$ by simply adding a new primitive predicate T to $\mathfrak{M}$ and stipulating that T be satisfied by all and only the true sentences

of $\mathfrak{M}$. However $\mathfrak{M}'$ is formed, Tarski's theorem naturally applies to it as well, and so $\mathfrak{M}'$ does not have its own truth predicate, though a richer language $\mathfrak{M}''$ may have a truth predicate for $\mathfrak{M}'$. The process of forming $\mathfrak{M}'$ from $\mathfrak{M}$ may be iterated, yielding a sequence $\mathfrak{M}$, $\mathfrak{M}'$, $\mathfrak{M}''$, ... of languages, each of which lacks its own truth predicate but has a truth predicate for the previous language.

One way this process could be carried out is as follows. Let $\mathcal{L}$ be any uninterpreted language, and let $\{T_1, T_2, \ldots\}$ be a collection of distinct unary predicates not contained in $\mathcal{L}$; let $\mathcal{L}^* = \mathcal{L} \cup \{T_1, T_2, \ldots\}$. Let $\mathfrak{M}$ be any classical model for $\mathcal{L}$ whose domain includes all the sentences of $\mathcal{L}^*$. Now let $\mathfrak{M}_0 = \mathfrak{M}$, and let $\mathfrak{M}_{n+1}$ be the expansion of $\mathfrak{M}_n$ to $\mathcal{L} \cup \{T_1, \ldots, T_{n+1}\}$ in which T_{n+1}'s extension is the set of sentences true in $\mathfrak{M}_n$. The resulting sequence $\mathfrak{M}_0, \mathfrak{M}_1, \ldots$ is called the *Tarskian hierarchy of languages* for $\mathfrak{M}$. We may also form a new model $\mathfrak{M}_\omega$, defined to be the expansion of $\mathfrak{M}$ to $\mathcal{L}^*$ in which each T_n has the same extension that it has in $\mathfrak{M}_n$; whether one deals with the sequence $\mathfrak{M}_0$, $\mathfrak{M}_1$, ... or the single language $\mathfrak{M}_\omega$ is largely a matter of taste. And of course $\mathfrak{M}_\omega$ itself may be further expanded by adding a new primitive truth predicate for $\mathfrak{M}_\omega$; but we won't be concerned here with extensions of the Tarskian hierarchy beyond its finite levels.[1]

The truth schema ''A' is true iff A' is now replaced by a sequence of restricted truth schemas; for each i, the following holds in $\mathfrak{M}_\omega$:

$$T_i(\text{`}\varphi\text{'}) \leftrightarrow \varphi, \text{where } \varphi \text{ is a sentence of } \mathcal{L} \cup \{T_1, \ldots T_{i-1}\}.$$

[1]There doesn't seem to be any canonical source in the theories of truth literature for the Tarskian hierarchy; it's part of the subject's folklore. Different versions of the hierarchy can be found, and no one version really deserves to be called the "correct" version. One thing that should be noted about the version just presented is that the usual formation rules of the first-order predicate calculus apply, so that sentences like $T_1(\text{`}T_2(\text{`}\varphi\text{'})\text{'})$ are well-formed, even though they are not true in any of the models $\mathfrak{M}_n$. Thus, while it is sometimes stated (e.g., in [**Bur79**, p. 85]) that the Tarskian approach to the paradoxes involves restricting the formation rules, this is not the case for every version of the approach. Indeed, no such restriction is to be found in [**Tar35**]: there, the predicate Tr is always a predicate of individuals, and $Tr(t)$ is always well formed, provided t is an individual term, e.g., a quotation name of a sentence, even if that sentence happens to contain the predicate Tr. It seems to me that remarks like Burge's involve confusing Tarski's and Russell's approaches.

We also have the following for each i and each sentence φ of $\mathcal{L}^*$:

$$\big[T_i(\text{`}\varphi\text{'}) \vee T_i(\text{`}\neg\varphi\text{'})\big] \to \big[T_i(\text{`}\varphi\text{'}) \leftrightarrow \varphi\big] \tag{7}$$

Notice that (7) is equivalent to the schema

$$T_i(\text{`}\varphi\text{'}) \to \varphi. \tag{8}$$

(Obviously (7) implies (8), so assume that (8) holds for all φ and suppose $T_i(\text{`}\varphi\text{'}) \vee T_i(\text{`}\neg\varphi\text{'})$. If $T_i(\text{`}\varphi\text{'})$ then $\varphi$ by (8), so assume $\varphi$ to show $T_i(\text{`}\varphi\text{'})$. If $\neg T_i(\text{`}\varphi\text{'})$ then $T_i(\text{`}\neg\varphi\text{'})$ by hypothesis, which implies $\neg\varphi$ by (8) with $\neg\varphi$ in place of φ; so $T_i(\text{`}\varphi\text{'})$.)

It is important to notice that the i in T_i is not a variable and hence cannot be bound by a quantifier. Or to put it another way, T is not a binary relation symbol. This is no accident, since the relation $\{(\varphi, i) : T_i(\text{`}\varphi\text{'}) \text{ holds in } \mathfrak{M}_\omega\}$ is easily seen to be indefinable in $\mathfrak{M}_\omega$: for if some formula $\tau(x, y)$ defined it, then $\exists y\, \tau(x, y)$ would be a truth predicate for $\mathfrak{M}_\omega$, in violation of Tarski's indefinability theorem.

In fact, we can generalize this last result.

Definition. A classical model $\mathfrak{M}$ for a language $\mathcal{L}$ is a *generalized Tarskian language* if (a) $\mathfrak{M}$ contains the sentences of $\mathcal{L}_M$, (b) substitution is definable in $\mathfrak{M}$, and there is a set $I \subseteq M$ and a family $\{\tau_i(x) : i \in I\}$ of formulas of $\mathcal{L}_M$ such that (c) $\tau_i(\text{`}\varphi\text{'}) \to \varphi$ holds in $\mathfrak{M}$ for each sentence $\varphi$ of $\mathcal{L}_M$, and (d) for each $\varphi$ there is an $i \in I$ such that $\mathfrak{M} \models \tau_i(\text{`}\varphi\text{'}) \vee \tau_i(\text{`}\neg\varphi\text{'})$.

Theorem 1. *If $\mathfrak{M}$ is a generalized Tarskian language, then the relation $\{(x, i) : \mathfrak{M} \models \tau_i(x)\}$ is not definable in $\mathfrak{M}$.*

Proof. Suppose $\psi(x, y)$ defined that relation. Then the formula $\exists y\, \psi(x, y)$ defines the set that consists of the true sentences of $\mathcal{L}_M$ and, perhaps, some nonsentences. But this is impossible, since substitution is definable in $\mathfrak{M}$ and therefore some sentence λ of $\mathcal{L}_M$ says $\exists y\, \psi(\text{`}\lambda\text{'}, y)$. □

One might (though Tarski himself did not) take the Tarskian hierarchy to be a model of natural language. That is, the English word 'true' (on this view) is ambiguous, and in fact has infinitely many senses, which we may write 'true_1', 'true_2', etc. Once 'true' is completely disambiguated by the addition of subscripts, 'true_n' is a truth predicate not for English as a whole, but for that fragment of English in which 'true_k' does not occur for $k \geq n$: that is, ''A' is true_n iff A' only holds when A belongs to this fragment of English.

The Liar is then handled neatly. A sentence $S =$ 'S is not true' must be construed as 'S is not true_n' for some n; then (1)–(3) become (for some k)

($1^\dagger$) S is not true_n

($2^\dagger$) 'S is not true_n' is true_k

($3^\dagger$) S is true_k

If k is greater than n, the argument is valid but harmless, since ($1^\dagger$) is perfectly consistent with ($3^\dagger$). On the other hand, if $k \leq n$ then the argument is invalid, since the move from ($1^\dagger$) to ($2^\dagger$) is not justified by the restricted truth schema for 'true_k'.

Several objections have been raised against this proposal. One is that 'true' simply does not seem ambiguous, at least if by that one means that it has different meanings in different contexts. Another objection concerns how the level of a given use of 'true' is determined. A natural assumption is that it goes some like this. If S is a nonsemantic sentence, i.e., one without any occurrences of 'true', then 'S is true' is to be rendered 'S is true_1'. If S contains (upon disambiguation) at least one occurrence of 'true_1' but no occurrences of 'true_n' for $n > 1$, then 'S is true' is rendered 'S is true_2'; and so on. On this approach, the level assigned to an occurrence of 'true' depends on the form of the sentence in which it occurs. The trouble is that this simply

won't do for many common uses of the truth predicate. Consider, for example, the sentence S = 'Every sentence on the front page of today's New York Times is true'. For S to have its intended sense, the subscript on the occurrence of 'true' in S must be higher than that assigned to any occurrences of 'true' on the front page of the relevant issue of the New York Times and what this is is an empirical matter, not determined by the form of S. Notice that it may also be beyond the ken of an utterer of S; so even if the foregoing sketched were modified to take the speaker's intentions into account, that by itself would not be adequate.

These problems might be avoided by a subtler theory that still takes the Tarskian hierarchy to be, in some sense, a model of natural language. An objection that is not easily met in this way was raised by Kripke. Suppose Dean says

(9) All of Nixon's utterances about Watergate are untrue.

In order to capture the intended sense of Dean's remark, the subscript on his utterance of 'true' must be greater than that on any of Nixon's utterances of 'true'. However, suppose that among the things Nixon says about Watergate is

(10) Everything Dean says about Watergate is untrue.

By the same reasoning, the subscript on Nixon's utterance of 'true' must be greater than that on any of Dean's utterances. But of course it is impossible to satisfy this condition for (9) and (10) simultaneously. Yet (9) and (10) are perfectly intelligible statements, and intuitively they could have straightforward and unproblematic truth values, under the right circumstances–for example, if everything besides (10) that Nixon says about Watergate is false, but Dean says at least one true thing about Watergate other than (9), then (10) is false and (9) is true.

Finally, and most importantly for my purposes, the Tarskian account of natural language is arguably self-defeating. To see this, let's first look at a simpleminded version of the account. Suppose truth is regarded as a relation between sentences and

positive integers, so that a sentence is true_n just in case it bears the truth relation to n. Then consider

(11) (11) is not true_n for any n.

First we show that (11) is not true_n for any n. For suppose (11) is true_k. By (the natural language analogue of) (8) we see that (11) is not true_n for any n, so in particular (11) is not true_k, contradiction. It follows that (11) is not true_k, and since k is arbitrary, we may conclude

(12) (11) is not true_n for any n.

But now we feel entitled to say that our conclusion itself is true; but something can only be true by being true_k for some k, so our conclusion must take the form

(13) '(11) is not true_n for any n' is true_k

which implies

(14) (11) is true_k

which contradicts (13). In short, if truth is construed as a relation, then the strengthened Liar problem is still with us.

It appears, then, that any viable version of the Tarskian account will not treat truth as a relation, and in particular will not allow the 'n' in 'true_n' to be bound by a quantifier. At the same time, in stating the Tarskian account one does make certain general claims about all the levels of the hierarchy–for example, that a sentence is true_n or false_n just in case it has no occurrences of 'true_k' for $k \geq n$. If this claim is to be stated in the language to which the Tarskian account applies, then it can't be stated in the obvious way, i.e., as a universal generalization with respect to 'n', and it is up to the proponent of the Tarskian account to tell us just how it is to be stated.

It's hard to say anything definite in the absence of a more specific account, but there is a distinct worry at this point. The worry is that however one tries to express the Tarskian account in a language to which it applies, the sort of generality that is used to express that account will enable the construction of Liar sentences analogous to (11).

Now this is no obstacle to the existence of languages that the Tarskian account is a correct account of. A Tarskian account of a language $\mathcal{L}$ might be advanced by speakers of an expressively richer metalanguage $\mathcal{L}'$, in which case the general claims which speakers of $\mathcal{L}'$ make about the levels of $\mathcal{L}$ might simply fail to be expressible in $\mathcal{L}$. The situation is different, however, when one advances a Tarskian account of one's own language, for in that case $\mathcal{L}$ and $\mathcal{L}'$ are one and the same, and a language cannot be expressively richer than itself. In other words, in trying to avoid the problem raised above when giving a Tarskian account of (say) English, we may wind up with a view that, by its own lights, is not statable in English. But since we are giving that account in English, this is an unacceptable situation.

So an adequate version of the Tarskian account must meet two constraints: it must be statable by its own lights, but the means by which it is stated must not allow for the construction of sentences like (11). (Or at least we need some explanation of why these constraints needn't be satisfied.) This is about all we can say in the absence of a particular account; in the next section I will look at Burge's account, which is basically a sophisticated version of the Tarskian account, and I will argue that it does not meet this pair of constraints. Analogous problems face every other instance of the hierarchy approach that I will consider: either it turns out to be unstatable by its own lights, or it involves concepts that can be used to form a new Liar sentence that the account can't account for. Or so I will argue.

It should be emphasized that this problem is essentially the same as the ordinary strengthened Liar problem for simple truth value gap accounts. There, as here, we have intuitively valid arguments with contradictory conclusions that the account

cannot explain away. And there, as here, we could try to avoid the problem by maintaining that the account is given in a metalanguage essentially richer than the object language it seeks to describe, though this would be unsatisfactory for the reasons given above. Every hierarchy account that I am familiar with prides itself on avoiding the strengthened Liar, and cites its (alleged) ability to do so as a major source of motivation; if it can be shown that such an account suffers from essentially the same problem in a different guise, then this would reveal a major shortcoming by the account's own lights.

2. Burge

Tyler Burge works out a theory of truth in his [**Bur79**] and [**Bur82**] that in many ways resembles the Tarskian hierarchy but avoids many of its problems. Like the Tarskian account, Burge's account assumes that natural language truth predicates are implicitly subscripted in some sense. This is not because 'true' is ambiguous, however, but rather because it is indexical. Moreover, what subscript a given token of 'true' receives does not depend on the form of the sentence in which it occurs, and is not assumed to be known by the speaker. A sentence containing 'true$_i$' is also no longer ineligible to be true$_i$, and so the account avoids the Nixon-Dean problem. Finally, Burge attempts to show that the account is statable in the very language it applies to.

2.1. The Account. The account consists of a "formal" part, which presents a formalized language with a hierarchy of truth predicates, and a "pragmatic" part, which interprets it. The formal language is more or less as follows. (See the appendix at the end of this chapter for the precise relation between what I'm about to describe and what Burge presents.) Let $\mathcal{L}$ be an (uninterpreted) language, and let $T_1, T_2, \ldots$ be distinct unary predicates not in $\mathcal{L}$; let $\mathcal{L}^+ = \mathcal{L} \cup \{T_1, T_2, \ldots\}$. We begin with a classical model $\mathfrak{M}_0$ for $\mathcal{L}$, which represents the nonsemantic part of a natural language; we assume that the sentences of $\mathcal{L}^+$ belong to $\mathfrak{M}_0$'s domain. Using the strong Kleene valuation scheme, we then form the least fixed point $(\mathfrak{M}_0, (E, A))$ over $\mathfrak{M}_0$, using T_1 as the truth predicate. Next we "close off" the partial model $(\mathfrak{M}_0, (E, A))$, i.e., we let $\mathfrak{M}_1$ be the classical model $(\mathfrak{M}_0, E)$. We then iterate the process, so that in general $\mathfrak{M}_{n+1} = (\mathfrak{M}_n, E_n)$, where $(\mathfrak{M}_n, (E_n, A_n))$ is the least strong Kleene fixed point over $\mathfrak{M}_n$ with T_{n+1} used as the truth predicate. Finally we let $\mathfrak{M}_\omega$ be the "union" of the sequence $\mathfrak{M}_0$, $\mathfrak{M}_1$, ... of models: that is, $\mathfrak{M}_\omega$ is the classical expansion of $\mathfrak{M}$ to $\mathcal{L}^+$ in which each T_n is interpreted as it is in $\mathfrak{M}_n$, or equivalently, as it is in any $\mathfrak{M}_k$ that interprets it at all.

(A difference between the above and Burge's own presentation is that in the latter, the formal language is specified by giving a list of sentence schemas of $\mathcal{L}^+$. Thus, whereas we have described the formal language from the outside, in an essentially richer metalanguage, Burge describes it from the inside. (Actually, I'm not sure that Burge has really managed to describe his formal language from within—see the appendix for more on this.) The importance of this is that ultimately Burge must give his account of English *in* English, not in some richer metalanguage; and the sentence schemas by which the formal language is specified represent formal principles governing the variable extension of 'true' which, according to Burge, are statable in the very language they describe. The above model-theoretic specification of $\mathfrak{M}_\omega$, although a good deal more perspicuous than Burge's list of schemas, unfortunately leaves out this aspect of the account.)

Let us define the *level* of a sentence of $\mathcal{L}^+$ to be the greatest i such that T_i occurs in that sentence, and 0 if no T_i occurs in it. The main difference between $\mathfrak{M}_\omega$ and Tarski's hierarchy of formal languages is that the predicate T_i may apply to sentences of level i, as well as sentences of level less than i. In the Tarskian hierarchy as well as in $\mathfrak{M}_\omega$ the restricted truth schema

$$(\mathrm{T}^*) \qquad \big[T_i(\text{`}\varphi\text{'}) \vee T_i(\text{`}\neg\varphi\text{'})\big] \rightarrow \big[T_i(\text{`}\varphi\text{'}) \leftrightarrow \varphi\big]$$

holds, but whereas in the Tarskian hierarchy the antecedent of (T^*) holds iff φ is of level less than i, in $\mathfrak{M}_\omega$ its antecedent often holds even when φ is of level i (though never when φ is of level greater than i). Indeed, as the definition of $\mathfrak{M}_\omega$ reveals, each of the T_is behaves much like the truth predicates in the fixed points of Kripke's construction, except that the underlying logic is classical. Sentences for which the antecedent of (T^*) holds are called *rooted*$_i$, and sentences for which it doesn't hold are called *unrooted*$_i$ or *pathological*$_i$.

What has this to do with natural languages like English? For present purposes, the extension of the English truth predicate (in a given context) is best regarded as

a set of *possible tokens* of English sentences. Each possible token of 'true' is assigned a level or subscript by rules that we will consider in a moment; once all subscripts have been assigned, the formal part of Burge's construction serves as a model of the resulting subscripted version of English. Specifically, to determine the extension of a given token t of 'true', first find its subscript i, and then look at T_i's extension in $\mathfrak{M}_\omega$. To the extent that $\mathfrak{M}$ adequately models the nonsemantic part of English, T_i's extension will mirror t's extension, in the sense that a possible token t' will belong to t's extension just in case t''s formalization belongs to T_i's extension. (What t''s formalization is will in turn depend on the global assignment of subscripts, if t' contains any tokens of 'true'.) The assignment of subscripts may therefore be thought of as an assignment of extensions, though the subscript assigned to a token of 'true' determines its extension only after all the subscript assignments have been made.

These rules for assigning subscripts are the "pragmatic" part of Burge's account. Burge describes three of the rules, though he admits that there may be others. The first is *Justice*: for any level i, an assignment of subscripts should not be made so as to make one sentence rather than another pathological$_i$ without some reason. The second is *Verity*: subscripts are assigned to occurrences of 'true' "so as to maximize the applicability of truth schemas to sentences and minimize attributions of rootlessness"– that is, so as to make as few sentences as possible pathological. Finally, there is the principle of *Minimization*: an occurrence of 'true' is assigned the lowest subscript that is compatible with the other principles.

Minimization and Justice are straightforward enough, but Verity needs some clarification. I characterized it just now as the principle that attributions of pathologicality should be minimized. But pathologicality is relative to a level: a sentence may be pathological$_i$ but not pathological$_j$ for $j > i$. This suggests that the principle is that subscripts should be assigned in such a way that each sentence is pathological relative to as few levels as possible. This interpretation is supported by the characterization of Verity quoted above, namely that subscripts should be assigned so as

to maximize the applicability of truth schemas: for when Burge says that the truth schema $T_i(\text{`}\varphi\text{'}) \leftrightarrow \varphi$ is applicable to a given sentence, what he means is that that schema holds when that sentence is substituted for φ, and a necessary and sufficient condition for this is that the sentence be rooted$_i$.

Unfortunately, this interpretation doesn't square with the rest of Burge's discussion of Verity. For example, he writes:

> If paradox is to be avoided, the subscript on a truth predicate in a quantified sentence of the form '$\forall x\,(A(x) \rightarrow \neg T_i(x))$' must sometimes be higher than the subscript on truth predicates in sentences that satisfy A. For example, if someone said, "Everything Descartes said that does not concern mechanics was true", the subscript on 'true' would be high enough to interpret satisfactorily or give truth conditions to everything Descartes said that did not concern mechanics. [**Bur79**, p. 109]

If i is the highest subscript to occur on a truth predicate occurring in a sentence in A's extension, then Verity apparently requires that subscripts be assigned so that 'All As are true' be rendered 'All As are true$_{i+1}$' or possibly 'All As are true$_i$', but certainly not as 'All As are true$_k$' for any $k < i$. Yet the latter sentence is rooted$_j$ for all $j > k$, whereas neither of the others is rooted$_j$ for any $j < i$. So Verity as we construed it above never favors either of the first two sentences to the third as a rendering of 'All As are true', and indeed may favor the third over the other two.

The second sentence of the quoted passage, along with other things Burge says in the course of explaining Verity, gives a better indication what that principle really is. Verity requires that the 'true' in 'All As are true' receive a subscript high enough "to interpret satisfactorily or give truth conditions to" everything in the extension of A, i.e., a subscript i such that everything in A's extension is rooted$_i$. The general principle, then, seems to be something like this: when 'true' is used to ascribe truth to an utterance or utterances, that use of 'true' should be assigned a subscript i sufficiently high that those utterances are rooted$_i$ (once they themselves have been

assigned subscripts). It should be noted that Verity is defeasible, being overruled in certain cases by a speaker's intentions or by conventions of language.

2.2. The Account at Work. Now let's look at how Burge's account handles various problems. First of all, the strengthened Liar is handled essentially the way Tarski's account handles it. Let (S) be an utterance of the sentence '(S) is true'. We first argue that

(15) (S) is not true.

Once we assert (15), we go on to reason that

(16) '(S) is not true' is true

and

(17) (S) is true,

apparently contradicting ourselves. However, the pragmatic rules assign subscripts to these tokens of 'true' in such a way that the resulting sentences no longer contradict each other. The occurrences of 'true' in (S) and (15) receive the same subscript i– this is not a straightforward application of the rules we've mentioned, but receives an independent argument, which we need not go into here. The 'true' in (16) and (17) receive a subscript $k > i$: this is assured by Verity. Since (15) says of a given sentence token that it is not true_i, and (17) says of it that it is true_k, (15) and (17) are completely consistent with each other.

Burge believes that a change of some sort takes place between the initial utterance (S) and the subsequent utterance (15), even though they are subscripted in exactly the same way. He describes this as a change in the *impicatures* associated with these utterances. Suppose the 'true' in (S) is assigned the subscript i. When (S) is uttered, it is implicated that (S) "is to be evaluated with the truth_i schema", i.e., that the

truth$_i$ schema holds for (S), or in other words that '(S) is not true$_i$' is true$_i$ iff (S) is not true$_i$. This implicature is seen to be false (because inconsistent), or as Burge puts it, (S) is seen to "lack truth$_i$ conditions". As we have seen, whenever the truth$_i$ schema fails to apply to a sentence, that sentence is not true$_i$; and once we see that the truth$_i$ schema fails to apply to (S), we assert that (S) is not true$_i$, this time *without* the implicature that '(S) is not true$_i$' is true$_i$ iff (S) is not true$_i$.[2]

So Burge's account handles the strengthened Liar problem just as well as Tarski's; it also avoids some of the latter's vices. The English word 'true' is not ambiguous on Burge's account, as we have seen. Also, the subscripting of tokens of 'true' is determined not by the form of a sentence or by explicit intentions of a speaker, but by principles that take into account various factors of which a speaker may be unaware: the subscript on the 'true' in 'All of Descartes' remarks not concerning mechanics are true', for example, is determined in part by what remarks not Descartes made that did not concern mechanics.

Burge's account also has less difficulty with Nixon-Dean examples than does Tarski's. Recall that in the Nixon-Dean case, two speakers each attribute untruth to all of the other's assertions. On the Tarskian account of truth, when someone says something of the form 'All Fs are true' or 'All Fs are false', the only way to do justice to his intentions is to assign to his utterance of 'true' or 'false' a subscript higher than the subscripts assigned to occurrences of 'true' in the sentences that satisfy F: otherwise, 'All Fs are true' will come out automatically false and 'All Fs are false' will come out trivially true. This is not so on Burge's account: it is often possible to adequately capture the sense of 'All Fs are true' merely by assigning to

[2] "Implicature" doesn't seem to me to be quite the appropriate term for what Burge is describing here. He argues that it is an implicature in the sense of [**Gri75**], since it has what Burge takes to be the defining feature of implicatures: cancellability. But Grice certainly never defined an implicature to be a cancellable presumption; in fact, he never *defined* the notion at all. And while he did define 'conversational implicature', cancellability is by no means the central criterion for whether an implicature is conversational. In any case, it *is* a central feature of implicature that it is part of communication; and the presumption Burge is describing need not be in any way communicated, since the Liar reasoning he describes can perfectly well be performed privately. 'Background assumption' really would have been a better term for what Burge has in mind.

the occurrence of 'true' a subscript at least as high as the levels of the sentences in the extension of 'F'.

Consider the Nixon-Dean case in particular. Let i be the least level such that everything Nixon or Dean says about Watergate, with the exception of what they say about the untruth of each other's assertions, is rooted$_i$. Rendering their utterances about each other as 'Everything Nixon says about Watergate is untrue$_i$' and 'Everything Dean says about Watergate is untrue$_i$' appears to adequately capture what they meant to say. For intuitively, if Dean said at least one true thing about Watergate (other than his assertion about Nixon), then what Nixon said was straightforwardly false, in which case what Dean said was straightforwardly either true or false, depending on what else Nixon said; and our subscript assignment bears these intuitions out. Our intuitions are likewise borne out if Nixon said at least one true thing about Watergate. Only when their assertions are *intuitively* paradoxical does this subscripting construe Nixon and Dean as ascribing untruth$_i$ to pathological$_i$ utterances. (Our subscript assignment is determined by the pragmatic principles as follows: Justice determines that the occurrences of 'true' in both utterances receive the same subscript; Verity determines that it is at least i; and Minimization determines that it is no greater than i.)

2.3. Some Objections. One virtue of Burge's account that was cited above is that unlike the Tarskian account, it treats 'true' as having a fixed meaning. However, as we also saw, 'true' has a variable extension on this account, i.e., it is *indexical*; and it's not clear that this is much more plausible. I think we have a very strong tendency to use 'true' and related words as though they were univocal. In fact, evidence for this tendency can be found in Burge's own writing.

Burge characterizes 'true' as a word whose *extension* varies from context to context; but 'extension' is a semantic term whose behavior ought to be similar to that of 'true'. In particular, on any given occasion of use 'extension' ought to carry an implicit subscript (the extension$_i$ of a predicate being the set of things that satisfy$_i$

that predicate). It's clear, however, that the intended sense of Burge's remarks would not be preserved if for any particular i we replaced 'extension' by 'extension$_i$'. For example, it's clear that Burge takes the class of extensions of 'true' in all possible contexts to be isomorphic, when ordered by inclusion, to the set of subscripts ordered by $<$; but among the extension$_i$s of 'true' in all possible contexts there is a most inclusive one, namely the set of true$_i$ sentences. Perhaps there is a better way than I have given of understanding such talk of a predicate's "extension"; the point is that both Burge and the reader easily fall into using 'extension' as though it were univocal. Likewise, Burge speaks of a truth schema 'applying' to a sentence, apparently meaning that the relevant instance of the schema is *true*; again, the intended sense would not be preserved by replacing 'applies' with 'applies$_i$'.

This objection is by no means intended as a knockdown argument. It is intended to show that accepting Burge's account has a certain intuitive price, though this price may be worth paying in order to get a consistent understanding of truth. One might think that this problem is common to all versions of the hierarchy approach; it will turn out, however, that one can do much to avoid this problem within that approach.

A second objection concerns Nixon-Dean problems. While Burge's account handles Kripke's example admirably, there are some variations on that example that it handles less well. The trouble stems from the fact that one always has a way of raising the level assigned to one's use of 'true', namely by engaging in a certain kind of strengthened Liar reasoning. As Burge puts it:

> Having gone through the reasoning that leads to counting a pathological$_i$ sentence true$_k$, we can get ourselves into hot water again by adding, perversely, "But this very sentence isn't". We may regard ourselves as having intentionally and anaphorically taken over the context of use for 'true$_k$'. To evaluate our perverse afterthought, we need a new context. [**Bur79**, p. 106]

That is, after making an assertion of the form 'S is true', I can immediately say 'But this very sentence isn't', meaning not just that my new statement is not true, but that it is not what I just said the sentence S is, namely true_i, where i is the subscript assigned to my first utterance. Thus, when I make my second utterance my intentions force the same subscript to be assigned to the occurrences of 'true' in both utterances. If I want to express the paradoxicality of the latter utterance, I may go on to assert 'My last utterance is not true', and I may even follow that up by asserting 'What I just asserted is true.' The occurrence of 'true' in my final utterance receives a subscript $k > i$.

Now suppose Sam makes the following series of utterances:

- What Fred says at noon is true.
- But this very utterance isn't.
- My last utterance is not true.
- What I just now asserted is true.

Suppose Fred makes a series of utterances of the same sentence types, but with 'Fred' changed to 'Sam'. Finally, suppose each makes his fourth utterance at noon exactly. I claim that there is no subscript assignment that gets this case intuitively right. Briefly, Sam's first utterance must be assigned a level at least as high as Fred's last utterance, for obvious reasons, and Fred's last utterance must be assigned a level strictly higher than that assigned to his first utterance, for the reasons just given; so Sam's first utterance gets assigned a level strictly higher than Fred's first utterance. But the same reasoning shows that Fred's first utterance must be assigned a level higher than that assigned to Sam's first utterance. But this is of course impossible.

Let's go over this more carefully. Remember that the whole point of the present account is to vindicate what we intuitively take to be the correct use of 'true'. The strengthened Liar reasoning both persons engage in seems to be *correct* reasoning, and treating it as such requires each person's fourth utterance to receive a higher subscript

than his second or third. But if Sam is correct in drawing his final conclusion, then surely Fred is correct in making his initial assertion, and vice versa. The intuitive correctness of these uses of 'true' also needs to be vindicated, and failure to do so would, I think, be at least as bad as failure to vindicate strengthened Liar reasoning. But endorsing Fred and Sam's initial assertions requires assigning each one a level at least as high as the levels assigned to the others' final assertions: for if Fred's last assertion is rendered 'What I just now asserted is true_k', then that assertion is only true_i when $i \geq k$.

The only remaining step of my argument is the claim that each person's second utterance is assigned the same level as his first. I suppose this could be coherently denied, even though doing so would run against Burge's own remarks. The assignment of subscripts to tokens is not, I think, something we have much of an intuitive grip on independently of what it entails about the correct use of 'true', so maybe we have no business insisting that the second utterance must be assigned the same level as the first. On the other hand, it does seem that I can have an intention to use 'true' with the same extension as a given previous use of that word. So the present account shares a vice with the Tarskian account, namely failure to do justice to our intentions.[3]

A third objection is as follows. If Burge's account is correct, then among the conventions of natural language are a complicated set of conventions governing the extension of 'true' in various contexts, and these are conventions we have learned. Presumably, too, these conventions existed before there was serious philosophical inquiry into the paradoxes. And it's fairly clear that their function is to provide

[3]Burge also offers an alternative construction that is more flexible than the one described here, in that it permits a certain kind of "looping" between levels. Call a sentence A *i-grounded* if it is grounded relative to the true_i sentences; for example, ''snow is white' is true_{17}' is i-grounded for all i, even when $i < 17$. The alternative construction allows i-grounded sentences to be true_i. (See the appendix for a more precise account.) One might think that this sort of looping gets around the above-mentioned problem; but it doesn't. First of all, when a sentence A ascribes truth to a sentence B, the occurrence of 'true' in A must still be assigned a subscript i high enough to make B i-grounded, if the intent of A is to be preserved. And second, when strengthened Liar reasoning begins by considering a sentence that self-ascribes untruth_i, the final conclusion of that reasoning (e.g., Sam's last sentence) is not i-grounded. These two facts prevent any adequate assignment of subscripts to Sams' and Fred's utterances.

natural language with a device of disquotation that is both consistent and useful: that is, their function is in part to avoid the paradoxes. All of this is surely implausible. Even though the rules governing the variable extension of 'true' are, according to Burge, operative all the time, they don't make any real impact on our use of 'true' outside of highly exceptional situations such as strengthened Liar reasoning, so it's quite mysterious how ordinary speakers could ever have mastered them. It's equally mysterious how such a sophisticated and fine-tuned device for avoiding paradoxes could have arisen in the absence of any awareness of the paradoxes.

In this respect, Burge's account differs from the partiality account of the last chapter — there, a story was told about how a partially defined truth predicate arises in a language for reasons having nothing to do with the paradoxes and without any awareness of the paradoxes on the part of the language's speakers. Likewise, we will soon encounter hierarchy accounts in which the contextually determined element that is present in truth ascriptions is a special case of a more general sort of context sensitivity that has nothing to do with the paradoxes. What Burge's account cries out for is some way of understanding the indexicality of 'true' in a similar fashion.

Finally, there is the strengthened Liar/unstatability problem, which we started to discuss in the first section of this chapter. Just as before, we see that if we regard truth as a two-place relation, then a strengthened Liar problem arises involving a sentence $S =$ 'S is not true$_i$ for any i'. But truth is not regarded as a two-place relation on Burge's account. Burge is explicit about this: "Attempts to produce a 'Super Liar' parasitic on our symbolism tend to betray a misunderstanding of the point of our account. For example, one might suggest a sentence like (a), '(a) is not true at any level'. But this is not an English reading of any sentence in our formalization." [**Bur79**, p. 108]

This leads us to ask, as we did about the Tarskian hierarchy, how general statements about arbitrary levels of the hierarchy are to be understood. Burge lays out several general principles governing truth$_i$ for various i, e.g., if $i < j$ then everything

that is true_i is true_j. These general principles appear to be stated in English; but how can they be, if every use of 'true' in English is associated with some particular level of the hierarchy?

When I said earlier that every token of 'true' receives a numerical index, I was simplifying a bit. According to Burge, this holds for most, but not all, uses of 'true', namely the *indexical* ones. Some uses of 'true' are *schematic*: that is, some possible sentence tokens containing 'true' are to be represented by sentence *schemas* of the formal language, in which the subscripts on 'true' are not numerals but schematic letters. Let us call such sentence tokens *schematic sentence tokens*. Schematic letters are not variables, in that they cannot be bound by quantifiers. On the other hand, schemas can be used to make generalizations about the whole hierarchy, so schematic letters should be thought of as standing for arbitrary levels. To assert a schema is, in effect, to assert all of its instances, i.e., to assert everything that results from substituting numerals for its schematic letters.[4]

Since schematic letters are not variables, we avoid at least one version of the strengthened Liar problem. However, there is also a schematic version (which Burge doesn't consider) which, while it is not in itself a problem for the account, is worth examining. Consider the following:

($) The schema marked with a ($) is not true_i.

There is no apparent obstacle to regarding the 'i' in ($) as a schematic letter, and hence to regarding the string of letters marked with a ($) as a schema. And since the extension of 'true_i' always consists of sentences rather than schemas, we would

[4]Burge offers a somewhat different characterization of the schematic: "In the formal principles [i.e., the axioms characterizing the formal language], the subscripts marking contexts of use stand open, ready to be filled in as the occasions arise." [**Bur79**, p. 107] This "filling in" is presumably the inferring of a particular instance from a general schematic principle. But it's not just *instances* of schematic principles that one asserts: the principles themselves can be asserted, and Burge's characterization of such principles leaves this fact out. Moreover, his characterization gives the impression that, being "open" and "ready to be filled in," schematic principles somehow say less than any one of their instances, whereas of course they say more.

apparently be justified in asserting that the schema marked with a (\$) is not true_i, where i is arbitrary, i.e., schematic. Thus we may draw the schematic conclusion

(\$) The schema marked with a (\$) is not true_i.

We are therefore led to assert the schema (\$), but for each n we are forced to conclude that what we have asserted is not true_n. We would like to say that the schema we have asserted is true, but we can't do so by means of an indexical use of 'true'. Is there some other way to do it?

For Burge, schemas can indeed be true, and we may assume that once one has assertively uttered a schema, one is licensed to infer that it is true. But the claim that a given schema is true is not the same as the claim that it is true_i for any particular i; rather, it is itself a schematic assertion. For example, to say that the schema '$\forall x$ (if x is true_i then x is true_{i+1})' is true is simply to assert the schema ''$\forall x$ (if x is true_i then x is true_{i+1})' is true_{i+2}'. The quotes here are to be understood as semiquotes: that is, a typical instance of ''$\forall x$ (if x is true_i then x is true_{i+1})' is true_{i+2}' is ''$\forall x$ (if x is true_{17} then x is true_{18})' is true_{19}', not ''$\forall x$ (if x is true_i then x is true_{i+1})' is true_{19}'.

Here, though, we begin to run into trouble. There is no evident way of rendering 'What Mary said yesterday is true' as a sentence or schema of the formal language when 'what Mary said yesterday' denotes a schematic utterance, even though 'What Mary said yesterday is true' seems to make perfect sense and intuitively might even be true. It can't be rendered as a *sentence* of the form 'What Mary said yesterday is true_i', for such a sentence is never true (at any level). Nor can it be rendered as the *schema* 'What Mary said yesterday is true_i', for to assert such a schema is to assert all of its instances, and we have just seen that each of its instances must be rejected. And if what Mary said yesterday is (rendered formally as) the schema 'Everything is either true_i or not true_i' (for example), then the schema ''Everything is either true_i or not true_i' is true_{i+1}' is by no means a correct translation of 'What Mary said yesterday is true', since it is an empirical matter what Mary said yesterday.. Similar

remarks apply to statements of the form 'All Fs are true' when the extension of 'F' includes some schematic utterances.

There is also no evident way to construe 'What Mary said yesterday is not true'. Here the difficulty is perhaps even greater. The last paragraph shows, in effect, that the formal language lacks a genuine truth *predicate* for schemas, having only a truth *operator*; in the case of untruth, we seem not even to have an operator. Negation is not such an operator, since the negation $\neg\Phi$ of a schema Φ says in effect that *every* instance of Φ is untrue, whereas to say that Φ is untrue is to say that *some* instance is untrue. In any case, there is no uniform way to say in the formal language that a schema is untrue, i.e., no formula schema $\Phi(x)$ such that, if t is a term denoting a schema Ψ, then $\Phi(t)$ says that Ψ is not true: to see this it suffices to consider a term t that denotes $\Phi(t)$.

Perhaps we could fix this by enlarging the formal language somehow. For instance, we might add a genuine truth predicate 'TRUE' for schemas, so that if Φ is a schema, 'TRUE(Φ)' is a *sentence*. Of course, we would like to be able to say that a sentence of the enlarged language is true. 'TRUE' might do double duty here, serving also as a truth predicate for sentences containing 'TRUE', or we might introduce yet another predicate 'true_ω' for this purpose. Either way, we will be able to construct new Liar sentences, which will require still yet another predicate '$\text{true}_{\omega+1}$' if the strengthened Liar is to be avoided. And this leaves us right back where we started. We have simply added a few new levels to the hierarchy, which clearly accomplishes nothing. Since the discussion so far has not once adverted to the fact that the hierarchy has order type ω, making the hierarchy taller cannot possibly rectify any of the problems we've raised.

Perhaps, then, we should reconsider the assumption that a schematic utterance can be true. Russell, whose typical ambiguity[5] bears a nonaccidental resemblance to Burge's schemas, thought that 'true' did not properly apply to schemas (in his

[5]as put forth in [**Rus08**]

terminology, formulas that contain "real" variables)–indeed, he thought a schema did not even express a single proposition. This position is hard to maintain, though, when such statements as 'Every sentence is true or not true' are understood schematically. On the one hand, that statement seems more general than any one sentence of the formal language could be; on the other, anyone willing to assert it would have no hesitation in calling it true. If a schema is explicitly presented as such, then perhaps a case could be made that what has been presented is a *rule* rather than a *statement*, and that to call it true would be a category mistake. Even here, though, it seems to me that I am entitled to say, of a schema I accept, that all of its instances are true, and that the latter statement, which clearly cannot be translated as a sentence of the formal language, is itself true.

For all of these reasons, I claim that Burge's discussion does not yield a plausible account of how his theory of truth can be simultaneously true of and statable in a natural language. If no such account can be given, then I think we must conclude that the theory is self-defeating.

3. Parsons

A rather different version of the hierarchy approach can be found in Charles Parsons' [**Par74**]. There he maintains that a truth ascription can be used to say different things on different occasions, but not because of the meaning of 'true'; the variability is due instead to the nature of natural language quantification. He takes truth to apply primarily to *propositions*; for a sentence to be true (on a given occasion) is for it to express a true proposition (on that occasion). Since a natural language quantifier has different ranges on different occasions, a sentence 'S expresses a true proposition' says different things on different occasions. Parsons exploits this variability in much the way Burge exploits the truth predicate's alleged indexicality.

As applied to propositions, 'true' is univocal: it is not indexical or ambiguous and has no hidden parameters. Also, as with most hierarchy accounts, the background logic is assumed to be entirely classical. The disquotational schema takes the following form:

(18) $$\forall x \left[\text{`}S\text{'} \text{ expresses } x \rightarrow (x \text{ is true } \leftrightarrow S)\right].$$

Now consider

(19) (19) does not express a true proposition.

Using (18), we may reason as follows. If (19) expresses a true proposition (p, say), then by (18), p is true iff (19) does not express a true proposition; it follows that (19) does not express a true proposition. (In fact, (19) does not express a proposition at all: we have just seen that (19) does not express a true proposition, so suppose (19) expresses an untrue proposition q; then by (18) we see that (19) expresses a true proposition, contradicting our earlier result.) Having reached this conclusion, we assert

(20) (19) does not express a true proposition.

But now we feel entitled to conclude that

(21) '(19) does not express a true proposition' expresses a true proposition

and, since (19) = '(19) does not express a true proposition', that

(22) (19) expresses a true proposition,

apparently contradicting (20).

The reason this is not a genuine contradiction is that there has been a change between (20) and (21) in the contextually determined domain of quantification. (19) does express a proposition p, but the quantifiers in (19) and (20) range over a domain of discourse D that does not include p. The quantifiers in (21) and (22), on the other hand, range over a larger domain D', which does include p. Thus (20) says that (19) does not express a true proposition that belongs to D, whereas (22) says that (19) does express a true proposition that belongs to D'. Thus what (20) and (22) say are mutually consistent, and indeed both are true.

In broad outline, this is Parsons' account of truth and the strengthened Liar; there are also some details and qualifications worth mentioning. To begin with, we have been assuming (and Parsons also assumes) that it is sentence types, rather than utterances, that express propositions. This is a simplifying assumption: we are all well aware that a sentence can express different propositions in different contexts, but we have been ignoring this fact in order to avoid irrelevant complications. It should be noted, however, that Parsons' account actually *requires* a single sentence type to express different propositions in different contexts, since what a sentence says depends on what domain its quantifiers range over, which in turn is a context-dependent matter. Strictly speaking, (18) ought to be rephrased so as to reflect this. And the modification cannot be something as trivial as

(18′) For all x and t, if t is a token of 'S' and t expresses x, then x is true iff S.

For one thing, (18′) implies that if two tokens of 'S' express propositions x and y, then x is true iff y is true; but clearly this is not always the case. Moreover, a token t of 'S' might have been uttered in a context relevantly different from that in which (18′) itself is uttered, in which case the consequent 'x is true iff S' may fail. On the other hand, certain special cases of (18′) are acceptable: if t is a token of 'S', then we should be willing to assert

(18″) For all x, if t expresses x then x is true iff S

in any context relevantly similar to that in which t itself is uttered (which for our purposes means that the contexts specify the same domain of quantification). I will not attempt a general reformulation of (18). Though explicitly rephrasing (18) may be difficult, (18)'s intuitive sense is fairly clear; and while it would be nice to have something more explicit, this intuitive understanding will generally prove sufficient for our purposes.

In light of all this, (19)–(22) may be reanalyzed as follows. First, each of '(19)', '(20)', etc. should be understood as referring to a particular utterance of the displayed sentence type, not to the type itself. The argument that (19) fails to express a true proposition remains sound, provided it takes place in a context relevantly similar to that in which (19) was uttered. (The argument now uses (18″) in place of (18).) The conclusion, (20), is a token of the same type as (19), and since the contexts are relevantly similar, (19) and (20) express the same proposition. (21) should, of course, be rewritten so that truth is not applied to a sentence type, e.g., as '(20) expresses a true proposition'; since (19) and (20) express the same proposition, the conclusion '(19) expresses a true proposition', which is simply (22), is equally correct.

We can also show that

(23) (19)'s domain of quantification fails to include the proposition, if any, that (19) expresses;

moreover, (23) holds no matter how its quantifiers are understood. For if (19) expresses a proposition p, then p is true iff (19) does not express a true proposition in the range of (19)'s quantifiers. (Here we use our intuitive understanding of (18).) If p belongs to the range of (19)'s quantifiers, then p is true iff p is not true; so p does not belong to the range of (19)'s quantifiers. If we want to avoid the strengthened Liar problem, we had better say that the quantifiers in (21) and (22) have more inclusive ranges than the quantifier in (19), and therefore that (19)'s quantifier is restricted, i.e., fails to range over absolutely everything. And since it would seem that (19)'s quantifier could be as inclusive as we please–i.e., as inclusive as any natural language quantifier in any context–Parsons' account seems to be committed to the thesis that natural language quantification is always restricted quantification. Parsons himself seems to endorse this view, and I will return to it later.

It should also be mentioned that not every version of the Liar involves explicit quantification, even when it is phrased so that 'true' is used as a predicate of propositions. Indeed, if there is an ordinary version of the Liar, it is probably closer to

(24) This very proposition is not true

or

(25) What I'm saying right now is not true.

Indeed, unless 'the' is regarded as a quantifier, there is no explicit quantification in

(26) The proposition expressed by (26) is not true.

Parsons' framework allows for more than one way to handle these cases. I take it that Parsons would want to view the phrase 'this very proposition' as it occurs in an utterance of (24) as non-denoting. That may well be the end of the matter: saying

that 'this very proposition' is non-denoting, and that (24) does not say something true, does not seem to generate a strengthened Liar problem.

On the other hand, we might want to say something along the following lines. Although (24) does not actually contain any quantifiers, an utterance u of (24) does have associated with it a domain of discourse D, and if u expresses a proposition at all, that proposition falls outside D. This in turn renders 'this very proposition', as it occurs in u, non-denoting, or at least renders the proposition u expresses false (or at least untrue). In this case, we may go on to say truthfully 'That proposition is not true', where 'That proposition' denotes the proposition expressed by (24); our subsequent utterance, of course, is associated with a domain of discourse D' properly including D. Whether or not (24) should be explained in these terms, something like this would seem to be required for (26).

Finally, it should be mentioned that for Parsons, the account we have been describing is really just a special case of a more general type of account in terms of "schemes of interpretation." In particular, the move from (20) to (21), where the quantifiers go from having a less inclusive to having a more inclusive range, is a special case of moving from a "less comprehensive" to a "more comprehensive" scheme for interpreting one's own utterances and those of others. In effect, a scheme of interpretation can be thought of as a rule for determining which sentences are true and which are false (and hence which are neither). We may assume that fhe following holds on any shcme:

(27) If 'S' is true or 'S' is false, then 'S' is true iff S

(or rather some refinement of (27) that takes context-sensitivity into account), and that a scheme A is more comprehensive than a scheme B if the set of sentences made true or false by A properly includes the set of sentences made true or false by B. For natural language, Parsons offers no account of schemes of interpretation apart from domains of quantification, let alone an account of the Liar other than that in terms of

domains of quantification. For this reason, I will confine this discussion to the special case of domains of quantification.

Parsons' account has one clear advantage over Burge's. As we saw, the latter seems to ascribe to ordinary speakers of natural languages mastery of a complicated set of conventions whose only purpose is to handle the paradoxes. Parsons' account makes no such attribution. The only feature of natural language it adverts to is the flexibility of its quantifiers, and this is a feature we have every reason to believe natural languages possess. We also saw in the section on Burge that there is a very strong tendency to think of the truth predicate as univocal. Parsons' account, to its credit, respects this tendency: whatever contextual variability there may be in truth ascriptions has to do with the contextual variability of natural language quantifiers rather than with the truth predicate itself (at least in the latter's primary use as a predicate of propositions).

There are, however, some potential sources of trouble. It is not enough to avoid genuine contradiction *simply* to make quantifiers context sensitive; it is essential that they also never have completely unrestricted range. Even this is not quite enough; the quantifiers in (19) (for example) must be not only restricted, but must also exclude from their range the proposition that (19) expresses. This last requirement is counterintuitive, since one feels that (19) is being used, or at least *could* be used, to say something about the very proposition (19) expresses, which would force (19)'s quantifier to include that proposition in its range.

It might seem plausible that the proposition (19) expresses always falls outside the range of (19)'s quantifier if we think of the latter as fixed in advance, established earlier in the conversation. But sometimes the range of a quantifier changes in the course of a conversation, and (19) could be used, not with the intention of quantifying over some previously established domain of discourse, but with the intention of quantifying over a domain large enough to include whatever proposition (19) may express. (This is especially evident when the speaker is unaware that (19) is the very

utterance he is making.) If Parsons' account is right, such an intention can never come off.

If such a speaker's intentions are incapable of being fulfilled, then the natural view is that he expresses no proposition at al lin uttering (19). On Parsons' account, however, he does express a proposition, but one that doesn't quite satisfy his intentions. Parsons' account could of course be modified to allow that in certain cases utterances like (19) fail to say anything whatsoever. I will postpone detailed discussion of this to the section on Barwise and Etchemendy, who suggest an analogous move. For now, let's just notice that under such a modification, the account of certain Liar sentences would be very different from the account Parsons develops. For suppose

(28) (28) does not express a true proposition

is an utterance that expresses no proposition whatever on the modified Parsons account. The usual strengthened Liar reasoning seems as plausible here as it does with other Liar sentences: the account commits us to asserting

(29) (28) does not express a true proposition

from which we feel entitled to conclude that the utterance (29) expresses a true proposition. We cannot, however, consistently conclude from this that (28) expresses a true proposition after all, since we are assuming that (28) expresses no proposition whatsoever. Thus the explanation of Liar reasoning in terms of shifting domains of quantification is unavailable here, and some other explanation is required. (Later on I will argue that such an explanation would face substantial difficulties.)

Another worry is as follows. As I said before, Parsons' account explains strengthened Liar reasoning in terms of the well-known phenomenon of context-dependent natural language quantification; but if the context sensitivity imputed to (19)–(22) is an instance of that phenomenon, it's an atypical one. Normally, the range of a quantifier as used on a given occasion is known to, or at least knowable by, the speaker,

as well as the other conversational participants (at least when communication is successful): if I look in the fridge and say 'There's no beer', there is no difficulty on anyone's part in determining what my quantifiers range over. In cases like (19)–(22), however, one has *no idea* what the quantifiers range over (beyond the fact that the quantifiers in (21) and (22) have wider ranges than those in (19) and (20)). This at least raises doubts about whether what's going on in (19)–(22) is the same as what's going on with 'There's no beer' (though of course it is far from being a knockdown refutation).

In any case, there is a more serious difficulty, namely the strengthened Liar/ unstatability problem, which in this case takes the form of unstatability. Parsons is committed to the view that all natural language quantification is restricted quantification: that is, on any given occasion a natural language quantifier is always used to quantify over something less than all there is. (We saw earlier that this is a consequence of his approach to the strengthened Liar.) On the face of it, though, this is self-defeating: in saying that one never quantifies over everything, one thereby, in using the word 'everything', does just that.

Let (R) be the principle that all quantification is restricted quantification. Notice that I am not saying that (R) is incoherent in the sense that it could not have been true; what I am claiming is that (R) is self-defeating. If (R) is true, then it is unstatable; this would seem to be sufficient grounds for rejecting it.[6] But in rejecting (R), we are not saying that (R) couldn't have been the case; if (R) had been true, however, we wouldn't have been able to say so.

The most obvious way to try to get around this difficulty is to maintain that contrary to appearances, (R)'s quantifiers are themselves restricted (or more accurately,

[6]Another type of view that is unstatable by its own lights is irrealism about content; on a view of this variety, no one ever states anything at all. Philosophers disagree on whether these views' unstatability constitutes grounds for rejecting them. In any case, the considerations are very different for a view like Parsons', since Parsons presumably accept the connection, which content irrealists reject, between assertively uttering a sentence and stating that which the sentence expresses; it is thus harder for someone like Parsons to advocate an avowedly unstatable view than it is for a content irrealist.

that in any natural language utterance of (R), the quantifiers are restricted). But this won't help us at all. If an utterance u of (R) has its quantifiers restricted to some domain D, then what u says is that no quantifier token in D ever ranges over everything in D. Since there is presumably no problem quantifying over all *tokens*, it is safe to assume that D includes all quantifier tokens, including those in u. But then what u says implies that the quantifiers in u, which range over D, do not range over the whole of D! In general, an utterance u of (R) either has a domain that excludes the quantifiers in u itself, or says something contradictory.

While Parsons seems to be aware of this problem, it isn't wholly clear what he has to say about it. In a footnote he writes:

> ...[O]ne might interpret the quantifiers of this entire paper as ranging over some sufficiently large set and thus produce a discourse to which the analysis of the Liar paradox here given would not apply. The same remark could be make about general set-theoretic discourses.... In each case the 'proponent' could come back and extend his discourse so that it would cover the newly envisaged case.
>
> ...[T]he generality which such a discourse as this paper has which transcends any particular set as range of its quantifiers must lie in a sort of systematic ambiguity, in that indefinitely many such sets will do. But one cannot express wherein the systematic ambiguity lies except in language that is subject to a similar systematic ambiguity. [**Par74**, p. 28 n.]

Parsons makes it clear later that what he refers to here as "systematic ambiguity" is nothing other than the fact that quantifier domains vary contextually and are never universal. But this makes the last sentence of the quoted passage puzzling. As I have been arguing, a language that satisfies (R) is precisely the sort of language in which (R) *cannot* be expressed; and a language that is systematically ambiguous in the relevant sense certainly satisfies (R), so (R) cannot be expressed in a systematically

ambiguous language. By the same token, 'L is systematically ambiguous' itself would seem to be unstatable in L if L is subject to systematic ambiguity.

Maybe we should look for a different construal of 'systematically ambiguous'. In the passage just quoted, Parsons said "the generality which such a discourse as this paper has ... must lie in some sort of systematic ambiguity, in that indefinitely many such sets [i.e., domains of discourse] will do." This suggests that, even on a given occasion of use, a quantifier might have an indefinite range. The best way I know to make this clear is to regard systematically ambiguous utterances as *schematic*, in the following sense. Let S be some sentence whose quantifiers are not explicitly restricted, and let S/D be the result of restricting all of S's quantifiers to D (where 'D' here is an uninterpreted letter). Until now we have assumed that to say that S is systematically ambiguous is to say that on any given occasion, what S says is what S/D says for some *particular* value of 'D', the value depending on the occasion; but perhaps, even on a single occasion, S says that S/D holds for *all* values of 'D'. It's not hard to see how systematically ambiguous expressions might be used to say that (R) is true of some language; and while (R) as it stands is not true of a language with systematic ambiguity, it is at least not obvious that such a language would be incapable of expressing the relevant facts about itself, e.g., that it is subject to systematic ambiguity.

The trouble is that systematic ambiguity may now be used to construct new Liar sentences that Parsons' account can't handle. For ease of exposition, I am going to adopt a convention for making explicit whether a sentence is being used schematically; specifically, S/D is to mean what S means when the latter is used schematically. (Notice that what proposition S/D expresses is not context dependent in any way that is relevant to this discussion; we may therefore regard the *type* S/D as expressing a proposition.) Now consider the schematic sentence

(A) (A) does not express a true proposition in D.

Assume for contradiction that (A) expresses a true proposition in D^*, where D^* is some fixed but arbitrary domain. From this supposition, we may draw the (schematic) conclusion

(A) does not express a true proposition in D.

Since 'D' denotes an arbitrary domain, a special case of this conclusion is

(A) does not express a true proposition in D^*,

which contradicts our hypothesis. So (A) does not express a true proposition in D^*. But since D^* was arbitrary, we may now conclude (unconditionally) that

(C) (A) does not express a true proposition in D

(where again 'D' is used as a schematic letter). We now feel entitled to infer that (C) is true, i.e., expresses a true proposition. But (C) is exactly the same sentence as (A), and what proposition (C) expresses is not context dependent; so to conclude that (C) expresses a true proposition is just to conclude that (A) does. But this directly contradicts (C).

Systematically ambiguous utterances in the sense just outlined are similar to Burge's schematic utterances, and some of the moves Burge makes are available to Parsons at this point. In particular, Parsons could say that (a) most uses of quantifiers are context sensitive in the way indicated earlier on in this section, i.e., on a given occasion, a quantifier has some definite range, determined by context, whereas some uses are schematic, and (b) the statement that a given schematic utterance is true is itself schematic. The trouble with this is that it would be subject to the same kind of objection that was raised against the corresponding move in Burge's account. Moreover, in construing 'u is true' differently depending on whether or not u is schematic, we would be giving up an attractive feature of Parsons' account, viz.,

that 'true' has a primary use in which it is univocal and on which other uses are derivative.

We have looked at two kinds of languages: those of which (R) is true, which seem to be incapable of expressing (R); and those in which there are certain individual sentences which are, in effect, about absolutely everything, and in which there arise versions of the Liar to which Parsons' account does not apply. Indeed, we have here a general strategy for objecting to accounts in the spirit of Parsons. Such an account will either say that no sentence can be used to say something about absolutely everything, in which case there will probably be a problem about statability; or it will say that some sentences do possess unlimited generality, in which case it may well be possible to use this generality to construct Liar sentences to which the account does not apply. I think some of the appeal of Parsons' account comes from not clearly distinguishing these two cases.

I think there is little hope of avoiding the conclusion that (R) is unstatable if true. Might there be some way of coherently endorsing Parsons' view *despite* this? This is a notoriously difficult question, and I don't claim to be able to answer it with any definitude.

First, we might follow Russell who, rather than claiming that all quantification is restricted quantification, seems to regard the very term 'everything' as incoherent. Now this is not acceptable as it stands, since there are plenty of obviously coherent uses of that term; what Russell means is that 'everything', 'all', etc. are incoherent when they *purport* to quantify over everything. Thus, we might abandon (R) as self-defeating and endorse instead

(R′) Any expression that purports to mention absolutely everything is incoherent.

While (R) is pretty obviously self-defeating, it is not entirely clear whether (R′) is, since it is not entirely clear whether the phrase 'purports to mention absolutely everything' itself purports to mention absolutely everything. I am inclined to think it does, but I will not insist on this.

Assuming (R′) is coherent, then, is it enough for Parsons' purposes? I think it is enough for some of them, but not all. Given any particular utterance of the propositional Liar sentence (19), Parsons can claim that it expresses a true proposition that falls outside its own domain of quantification. Or, if its quantifier purports to be absolutely unrestricted, he can reject it as unintelligible on the basis of (R′). What Parsons *doesn't* seem to be able to do is to give any kind of *general* account of propositional Liar sentences. He certainly can't say that every such sentence expresses a true proposition that falls outside its own domain of quantification: for were he to say so, he would have to do so by means of a statement with some domain D of quantification, and that statement will simply not be true of those propositional Liars which themselves have domain D (or more generally a domain containing D).

There is also another limitation worth noting. As I just mentioned, when confronted with any particular propositional Liar sentence S, Parsons can offer a diagnosis of S, i.e., he can claim that S expresses a true proposition not belonging to the range of its own quantifiers. However, *Parsons* is not in a good position to say that he can always this: he cannot coherently claim that all such S can be successfully diagnosed. For to successfully diagnose S is to extend the range of one's quantifiers so that they include S's domain and also the proposition S expresses; so to say that S can be successfully diagnosed is, among other things, to say that S expresses a true proposition not belonging to the range of S's quantifiers. Thus, to claim that every propositional Liar sentence can be correctly diagnosed is tantamount to saying that every propositional Liar sentence expresses a true proposition that falls outside its own domain of quantification; and so if Parsons were to say that all such sentences admit of a successful diagnosis, he would be making precisely the sort of general claim that I just observed he cannot coherently make.

So even if (R′) itself coherent is, it leaves Parsons in the position of being able neither to coherently put forward a general account of propositional Liar sentences, nor even to claim that such an account can be given in every particular case. In other

words, if Parsons is not talking nonsense, then he has not given us a general account at all; he has at most explained a few particular propositional Liars, but has left it quite open that other such sentences will resist his kind of analysis. And the feeling that such an account could always be given for these other sentences is an illusion, as we have seen, since any attempt on Parsons' part to say that such an account could be given is self-defeating. I admit that, at this point at least, we don't have an account here that is subject to strengthened Liar difficulties or any obvious incoherencies; but this is because we don't have a general account at all.

Another possible way out is to maintain that (R), while not statable, is still *thinkable*. We might be able to believe (R), entertain (R), etc. without (R) being the conventional, linguistic meaning of any sentence in our language.

While I have no objection to the idea that some things could be thinkable but not statable, I don't think that idea helps in this case. To begin with, notice that (R) involves quantification over absolutely everything, so grasping (R) mentally would involve mentally quantifying over absolutely everything. If one can do this, then it would be implausible to deny that one can also grasp the thought that absolutely everything is F, or that not absolutely everything is F, provided F is a graspable property. There are two ways this can lead to trouble. First, notice that (R) had better be not only thinkable but *communicable*: this would seem to be a precondition for having any public discussion about (R) at all. (If Parsons' views turned out not to be communicable, then one wonders what he could be up to in writing his papers, or what the rest of us could be up to in discussing them.) In this case, even if (R) is not the linguistic meaning of any utterance, (R) may very well be the *speaker* meaning of some utterance. By the same token, the thought that everything is F is presumably also communicable and hence, it would seem, potentially the speaker meaning of some utterance (at least for suitable F–surely at least when Fness is expressible in public language). But in that case we have a curious situation: an expression of the form 'Everything is F' can be used to *mean* that absolutely everything is F, but cannot be

used to *say* that absolutely everything is F. While there is no outright contradiction here, this does seem implausible in the extreme.

Second, once we admit that it is thinkable that absolutely everything is F, we open ourselves up to the possibility of a completely *mental* version of the Liar, which Parsons' account can't accommodate. For in that case, I can have a thought T to the effect that T expresses absolutely no true proposition. I am then in a position to engage in strengthened Liar reasoning mentally. Parsons' strategy would then be inapplicable, since it involves widening the scope of one's quantifiers beyond what they were in the instance of the Liar with which one began: and one cannot, even in thought, quantify over more than there is.

4. Barwise and Etchemendy

The account of truth that Barwise and Etchemendy develop in [**BE87**] is much like Parsons' view, except that the role that domains of quantification played in that account is now played by the *situations* of situation semantics. A situation is, roughly, the portion of the world that a given statement is about, and is determined by the context of utterance as well as the conventions of language. Strengthened Liar reasoning is then explained in terms of a shift in the contextually determined situation. The account therefore shares a virtue with Parsons' account, namely, the contextually determined element present in truth ascriptions is a byproduct of a more general kind of context sensitivity which in and of itself has nothing to do with the paradoxes.

Another similarity to Parsons is that truth is taken to apply primarily to propositions rather than to sentences. There are two things that go into determining what proposition a sentence expresses on a given occasion of use: "demonstrative conventions," which pick out a situation s, and "descriptive conventions" of language, which pick out a type of situation T; the resulting proposition $\{s; T\}$ says that the situation s is of type T. (This talk of "descriptive" and "demonstrative" conventions seems to mean more or less that the type T is determined by the sentence's linguistic meaning, whereas the situation s is determined by the context of utterance.) Barwise and Etchemendy illustrate this view of propositions, which they call the "Austinian" view[7] and contrast with the more usual "Russellian" view, with a simple example:

> If the sentence "Claire has the ace of hearts" is used to describe a particular poker hand, then on the Austinian view the speaker has made a claim that the relevant situation is of the type in which Claire has the ace of hearts. Notice that such a claim could fail simply because Claire wasn't present, even if Claire had the ace of hearts in a card game

[7]The view is so-called because it is based on that of [**Aus50**].

> across town. By contrast, on the Russellian view the claim would be true. [**BE87**, p. 29]

Propositions are allowed to be self-referential (e.g., the Liar sentence expresses a self-referential proposition rather than failing to express a proposition). They are also modeled set-theoretically, and the most natural way to do this is to adopt a set theory that allows "circular" sets.

4.1. Non-Well-Founded Sets. These "circular" sets are more properly called *non-well-founded sets*: a non-well-founded set is simply a set x such that there is a sequence $x = x_1, x_2, \ldots$ of sets with $x_{n+1} \in x_n$ for all n. If $x \in x$, for example, then x is non-well-founded since we may take $x_n = x$ for all n. More generally, if $x \in x_1 \in \cdots \in x_n \in x$, then x is non-well-founded by similar reasoning.[8] Standard set theory (i.e., ZFC) has an axiom (Foundation) that says all sets are well-founded. However, it is consistent to replace this axiom with any of various axioms asserting the existence of non-well-founded sets. Barwise and Etchemendy take as their background set theory Peter Aczel's ZFC/AFA, which is obtained from ZFC by replacing Foundation by AFA, the "anti-Foundation axiom," which we will now explain.[9]

There is a natural way of representing sets graphically. For example, Figure 1 represents the set $3 = \{0, 1, 2\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$. Here, each dot d represents the

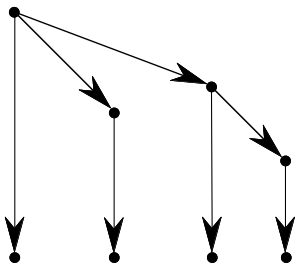

FIGURE 1

[8]A non-well-founded set needn't display this sort of direct or indirect self-membership; for example, let $x_1, x_2, \ldots$ be such that $x_n = (n, x_{n+1})$. So 'circular' is not an entirely appropriate term for these sets in general.

[9]Actually, their background set theory differs from ZFC/AFA in two respects: it allows urelements, and it has a global version of the axiom of choice, i.e., it posits the existence of a well-ordering of the entire universe of sets.

set of things represented by dots that d "sees," i.e., dots d' such that there is an arrow leading from d to d'. Since the dots at the bottom of the picture don't see anything, they all represent $\emptyset$; dots that see only dots at the bottom of the picture represent $\{\emptyset\}$; and so on. Every set can be represented by such a picture (in a sufficiently broad sense of 'picture'); the idea behind AFA is that the converse holds, i.e., every picture represents some set.

This is made precise by replacing talk of pictures with talk of graphs. A *graph* is a structure $G = (X, R)$, where R is a binary relation on X; X's elements are called *nodes*, and the ordered pairs in R are called *edges*. A *decoration* of a graph G is a function d whose domain is the set of G's nodes, and that satisfies $d(a) = \{d(b) : (a, b) \in R\}$ for each node a. The claim that every set is represented by a picture now becomes the claim that for every set x there is a graph G, a decoration d of G, and a node n of G such that $x = d(n)$. To see that this claim is true, let x be any set and let t be its transitive closure, i.e., the least transitive set of which x is an element; let $G = (t, R)$, where $R = \{(a, b) : a, b \in t \ \& \ b \in a\}$; and let d be the identity function restricted to t. Then d is a decoration of G, and of course $d(x) = x$.

AFA is the assertion that every graph has a unique decoration. Thus we see, for example, that AFA implies that there is a unique set $\Omega = \{\Omega\}$, obtained from the unique decoration either of the graphs in Figure 2. (To see that they give rise to the

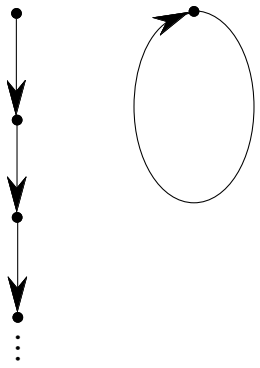

FIGURE 2

same set, let Ω be the set assigned to the node in the right-hand graph by its unique

decoration, and let $f(n) = \Omega$ for each node n of the left-hand graph. Clearly f is a decoration, and is therefore the left-hand graph's unique decoration.)

There are some tools that are useful when working in ZFC/AFA. In particular, there are two kinds of self-referential definition available to us. The first kind is made possible by a result called the *Solution Lemma*. Intuitively, the set Ω is the unique solution to the equation $x = \{x\}$; the Solution Lemma states that any such equation has a unique solution, and more generally, that any system of such equations has a unique simultaneous solution. Somewhat more formally, assume as given a class of "indeterminates"; the nature of the indeterminates is unimportant, though it may be helpful to think of them as urelements. If S is a set and f is a function that assigns a set to each indeterminate in S's transitive closure, let $f[S]$ be the result of "replacing" each indeterminate x by $f(x)$ throughout S. (If S is the set $(x, 1)$, for example, and x is an indeterminate, then $f[S]$ is the set $(f(x), 1)$.) If X is a set of indeterminates and $\{S_x : x \in X\}$ is any family of sets that are all built up from X, then the Solution Lemma states that there is a unique function s with domain X such that $s(x) = s[S_x]$ for all $x \in X$. For a more rigorous treatement of these matters see Chapter 1 of [**Acz88**].

The Solution Lemma allows for a natural set-theoretic modeling of self-referential propositions. A truth-teller proposition, for example, may be represented by the set $p = (Tr, p)$, where Tr is some arbitrary object chosen to represent truth: that is, p is the unique solution to the equation $x = (Tr, x)$. Using systems of two or more equations, we can represent systems of propositions that make reference to each other. For example, let p and q be the unique sets such that $p = (Tr, q)$ and $q = (Fa, p)$, Fa being an object that represents falsehood: that is, p and q are the unique solution to the pair of equations $x = (Tr, y)$, $y = (Fa, x)$. Then p and q represent a Liar cycle.

The other kind of self-referential definition available to us is the familiar *inductive* definition. In the presence of AFA, however, such definitions must be approached a bit delicately. Let Γ be a class operator (i.e., a function from classes to classes); Γ is said

to be *monotone* if $X \subseteq Y$ implies $\Gamma(X) \subseteq \Gamma(Y)$, and *set-based* if, whenever $x \in \Gamma(X)$, there is a *set* $y \subseteq X$ such that $x \in \Gamma(y)$; an operator that is both monotone and set-based is called *continuous*. A class X is called a *fixed point* of Γ if $X = \Gamma(X)$. We can prove facts about the existence of fixed points similar to those of Chapter 1 (though we can't simply apply those results here, since they apply only to a monotone function whose domain is a *set*, whereas Γ's domain is a superclass). Without assuming either Foundation or AFA we can show that every continuous operator has both a least and a greatest fixed point, i.e., fixed points X and Y such that $X \subseteq Z \subseteq Y$ for each fixed point Z. Specifically, set

$$\begin{aligned} X_0 &= \emptyset & Y_0 &= V \\ X_{\alpha+1} &= \Gamma(X_\alpha) & Y_{\alpha+1} &= \Gamma(Y_\alpha) \\ X_\lambda &= \bigcup_{\xi<\lambda} X_\xi & Y_\lambda &= \bigcap_{\xi<\lambda} Y_\xi \end{aligned}$$

where λ is a limit ordinal or ∞ and where V is the class of all sets. It's quite straightforward to show that X_∞ and Y_∞ are the desired least and greatest fixed points of Γ, respectively.

It often happens that an operator Γ that has a unique fixed point if Foundation holds has several fixed points if AFA holds. For example, we might informally offer the following as a definition of "hereditarily finite":

> A set is hereditarily finite iff it is finite and its elements are hereditarily finite.

What justifies this definition is the fact that, assuming Foundation, the continuous operator $\Gamma(X) = \{\text{finite subsets of } X\}$ has a unique fixed point, the class of hereditarily finite sets. But Γ has more than one fixed point if AFA holds: its least fixed point is the class HF_0 of well-founded hereditarily finite sets, and its greatest fixed point is the set HF_1 of sets represented by finite graphs. In what follows, we will frequently be interested in the greatest fixed point of a continuous operator.

The observation that each continuous operator has a greatest and least fixed point can be generalized. A class X is *sound* (w.r.t. Γ) if $X \subseteq \Gamma(X)$, and *closed* if $\Gamma(X) \subseteq X$; the proof that Γ has a greatest and least fixed point also shows that if X is sound then there is a least Γ-fixed point Y with $X \subseteq Y$, and if X is closed then there is a greatest Γ-fixed point Z with $Z \subseteq X$.

4.2. Russellian Propositions. Barwise and Etchemendy begin with a formal account of truth for "Russellian" propositions. The Russellian account is essentially a propositional version of Kripke's fixed point construction, and they ultimately reject it in favor of the "Austinian" account. Nonetheless, some facts about Russellian propositions will be needed in our discussion of the Austinian account; for this reason, and because the mathematical techniques will provide a useful background when we get to Austinian propositions, I will now give a summary of the Russellian account.

Russellian propositions are ordinary structured propositions. For simplicity, Barwise and Etchemendy only consider propositions that are about a card game between two people, Claire and Max; atomic Russellian propositions all belong to the following types:

- $[a\ H\ c]$ (a has c), where a is Claire or Max and c is a card
- $[a\ Bel\ p]$ (a believes p), where a is Claire or Max and p is a Russellian proposition
- $[Tr\, p]$ (p is true), where p is a Russellian proposition.

Each atomic proposition p has a negation $\overline{p}$. (We will generally write $[Fa\, p]$ instead of $[\overline{Tr\, p}]$.) If X is a set of Russellian propositions, then $[\bigwedge X]$ and $[\bigvee X]$ are Russellian propositions, called the *conjunction* and *disjunction* of X, respectively. The Russellian propositions are just those things that can be built up from atomic and negated atomic Russellian propositions by conjunction and disjunction.

Notice that arbitrary propositions occur as components of atomic propositions. This allows a proposition to be a component of itself, at least in principle, for nothing

we have said so far rules out the existence of, for example, a proposition $[Tr\, p]$ such that p is the proposition $[Tr\, p]$ itself. Of course, this is precisely what Barwise and Etchemendy want.

Formally, the class of Russellian propositions is defined as follows. First, choose some arbitrary objects to be Claire and Max and the playing cards. Next, choose some way of set-theoretically modeling the propositions $[a\ H\ c]$, $[\overline{a\ H\ c}]$, $[a\ Bel\ p]$, etc. in terms of a, c, p, etc.[10] Now given a class X, define $AtPROP(X)$ to be the class of all sets of the form $[a\ H\ c]$, $[a\ Bel\ x]$, and $[Tr\, x]$ and their negations, for $x \in X$. Define $\Gamma(X)$ to be the class of all sets that can be built up from the elements of $AtPROP(X)$ by conjunction and disjunction–that is, $\Gamma(X)$ is the the least fixed point of the continuous operator Δ that contains $AtPROP(X)$, where $\Delta(Y) = Y \cup \{[\bigwedge y] : y \subseteq Y\} \cup \{[\bigvee y] : y \subseteq Y\}$. Γ is clearly continuous, and as such has a least and a greatest fixed point; the class $PROP$ of Russellian propositions is defined to be Γ's greatest fixed point.

The choice of *greatest* fixed point assures that $PROP$ contains a rich supply of self-referential propositions. For example, let t be the unique set such that $t = [Tr\, t]$. Suppose for contradiction that t is not an element of $PROP$. $PROP \cup \{t\}$ is sound with respect to Γ: $PROP = \Gamma(PROP) \subseteq \Gamma(PROP \cup \{t\})$ by Γ's monotonicity, and $t \in AtPROP(PROP \cup \{t\}) \subseteq \Gamma(PROP \cup \{t\})$. Therefore there is a Γ-fixed point that extends $PROP \cup \{t\}$ and thus properly extends $PROP$; but this is impossible, since $PROP$ is Γ's *greatest* fixed point. More generally, this argument shows that $x \in PROP$ whenever $x \in \Gamma(PROP \cup \{x\})$.

(Notice that the definition of $PROP$ is somewhat roundabout: we first define a continuous operator Δ, then define another continuous operator Γ by $\Gamma(X) =$ least Δ-fixed point extending $AtPROP(X)$, and finally define $PROP$ to be Γ's greatest fixed point. Why not proceed more directly: say, define $\Phi(X) = AtPROP(X) \cup \{[\bigwedge x] :$

[10] We must assume that p belongs to the transitive closure of the atomic propositions $[a\ Bel\ p]$ and $[Tr\, p]$ and their negations in order to apply the Solution Lemma.

$x \subseteq X\} \cup \{[\bigvee x] : x \subseteq X\}$, and define *PROP* to be Φ's greatest fixed point? The answer is that doing so would allow into *PROP* such "contentless" propositions as $p = [\bigwedge\{p\}]$ or $q = [\bigvee\{q\}]$.)

Truth for Russellian propositions is relative to a *model*, which may be thought of as a possible world. Formally, a model is a class of atomic and negated atomic propositions that satisfies certain conditions.[11] A class $\mathfrak{M}$ of atomic and negated atomic propositions is said to be *coherent* if no atomic proposition and its negation belong to $\mathfrak{M}$. If $\mathfrak{M}$ is coherent then the relation $\mathfrak{M} \models p$ (p is made true by $\mathfrak{M}$), p an arbitrary proposition, is defined as follows:[12]

- If p is an atomic or negated atomic proposition, then $\mathfrak{M} \models p$ iff $p \in \mathfrak{M}$;
- $\mathfrak{M} \models [\bigwedge X]$ iff $\mathfrak{M} \models p$ for all $p \in X$;
- $\mathfrak{M} \models [\bigvee X]$ iff $\mathfrak{M} \models p$ for at least one $p \in X$.

That the relation $\models$ exists and is unique may not be wholly obvious, since the above has the form of a definition by induction on a well-founded relation, yet many of the sets involved are not well-founded. However, since each proposition is built up from the atomic propositions by conjunction and disjunction, no *new* non-well-foundedness is introduced at this stage. That is, the relation $R = \{(p, q) : p, q \in PROP$ and for some X, $p \in X$ and q is $[\bigvee X]$ or $[\bigwedge X]\}$ is well-founded, and its minimal elements are precisely the atomic and negated atomic propositions, so we can define $\models$ by induction on R.[13] Other inductive definitions given in this section should be understood similarly. For instance, the negation $\overline{p}$ of a proposition p is defined inductively as follows:

[11] Actually, Barwise and Etchemendy define a Russellian model to be a class of *states of affairs* that satisfy certain properties. However, there is a natural 1–1 correspondence between states of affairs and atomic and negated atomic propositions, and it does no harm to identify them.

[12] Here again Barwise and Etchemendy proceed slightly differently, first defining $\mathfrak{M} \models p$ for the special case in which $\mathfrak{M}$ is a set, then defining the general case by $\mathfrak{M} \models p$ iff $s \models p$ for some $s \subseteq \mathfrak{M}$. As far as I can tell, however, this detour is unnecessary: the definition given here is formally correct and equivalent to that given in *The Liar*.

[13] To see that R is well-founded, let R' be its ancestral and let S be the class of propositions q such that $R' \upharpoonright \{p : (p, q) \in R'\}$ is well-founded; then $R \upharpoonright S$ is well-founded. Now S is clearly a fixed point of Δ, and obviously $AtPROP(PROP) \subseteq S$, so $\Gamma(PROP) \subseteq S$; but $\Gamma(PROP) = PROP$. So $S = PROP$ and $R \upharpoonright S$ is simply R.

- $\overline{\overline{p}} = p$ when p is atomic;
- $[\overline{\bigwedge p}] = [\bigvee\{\overline{p} : p \in X\}]$;
- $[\overline{\bigvee p}] = [\bigwedge\{\overline{p} : p \in X\}]$.

A *weak model* is a coherent class $\mathfrak{M}$ of atomic and negated atomic propositions such that for all p,

- If $[Tr\,p] \in \mathfrak{M}$ then $\mathfrak{M} \models p$;
- If $[Fa\,p] \in \mathfrak{M}$ then not $\mathfrak{M} \models p$.

If we also have

- If $[Fa\,p] \in \mathfrak{M}$ then $\mathfrak{M} \models \overline{p}$

then $\mathfrak{M}$ is said to satisfy the *witnessing condition*. Weak models that satisfy the witnessing condition correspond to the sound points of Kripke's construction. It's quite easy to show that a proposition and its negation are never both made true by a weak model. A weak model is said to be *T-closed* if $\mathfrak{M} \models p$ just in case $\mathfrak{M} \models [Tr\,p]$, *F-closed* if $\mathfrak{M} \not\models p$ just in case $\mathfrak{M} \models [Fa\,p]$, and *N-closed* if $\mathfrak{M} \models \overline{p}$ just in case $\mathfrak{M} \models [Fa\,p]$. No weak model can be F-closed: for let $p = [Fa\,p]$ and suppose $\mathfrak{M}$ is F-closed, so that $\mathfrak{M} \not\models p$ iff $\mathfrak{M} \models [Fa\,p]$; but this is just to say that $\mathfrak{M} \not\models p$ iff $\mathfrak{M} \models p$, since $p = [Fa\,p]$. However, a weak model can be both T- and N-closed. A *model of the world* is defined to be a T- and N-closed weak model; in other words, a coherent class $\mathfrak{M}$ is a model of the world iff for all p, $\mathfrak{M} \models p$ iff $[Tr\,p] \in \mathfrak{M}$ and $\mathfrak{M} \models \overline{p}$ iff $[Fa\,p] \in M$. Models of the world correspond to the fixed points of Kripke's construction, and their existence is proved in basically the same way.

Notice that if $p = [Fa\,p]$ and $\mathfrak{M}$ is a model of the world, then neither $\mathfrak{M} \models p$ nor $\mathfrak{M} \models \overline{p}$, and so $\mathfrak{M} \not\models p$ but not $\mathfrak{M} \models [Fa\,p]$. Or as Barwise and Etchemendy put it, p is made false by $\mathfrak{M}$, but p's falsity is not a fact of $\mathfrak{M}$. If $\mathfrak{M}$ is taken to represent all the facts there are, then although p is false, it is not a fact that p is false. This is essentially the strengthened Liar problem, and it is Barwise and Etchemendy's principal complaint about the Russellian approach.

Finally, Barwise and Etchemendy also describe a formal language $\mathcal{L}$, and rigorously define 'formula φ expresses proposition p in context c', where p is a Russellian proposition. (Later they do the same thing for Austinian propositions.) The details are fairly straightforward and I won't bother to repeat them here. One thing that should be mentioned, though, is that $\mathcal{L}$ has a truth predicate, as well as a demonstrative '**this**' that denotes the proposition expressed by the formula in which it occurs. Thus, the sentence $\neg$**True**(**this**) expresses the proposition $p = [Fa\, p]$. The language also has demonstratives '$\mathbf{that}_1$', '$\mathbf{that}_2$', ..., whose referents are specified by the context c.

4.3. Austinian Propositions. We can now apply the techniques developed in the course of modeling Russellian propositions to the somewhat more complicated case of Austinian propositions. I said earlier that a situation is a portion of the world, but it would be more accurate to say that it is a collection of facts, or more generally of states of affairs. A type is, intuitively, a condition that a situation may or may not satisfy, and a proposition is a situation together with a type. To capture this formally, we begin by simultaneously defining the classes SOA, SIT, $AtTYPE$, and $PROP$ (of states of affairs, situations, atomic types, and propositions, respectively) to be the largest classes satisfying the following:

- Every $\sigma \in SOA$ is of one of the forms
  - $\langle H, a, c; i\rangle$,
  - $\langle Bel, a, p; i\rangle$,
  - $\langle Tr, p; i\rangle$,

  where $i = 0$ or 1, a is Max or Claire, c is a card, and $p \in PROP$;
- Every $s \in SIT$ is a subset of SOA;
- Every $p \in PROP$ is of the form $\{s; T\}$, with $s \in SIT$ and $T \in \Gamma(AtTYPE)$;
- Every $T \in AtTYPE$ is of the form $[\sigma]$, with $\sigma \in SOA$.

Γ here is the same as the operator Γ defined in the previous subsection. A *type* is an element of the class $TYPE = \Gamma(AtTYPE)$. As with Russellian propositions, $\langle H, a, c; i\rangle$ is built up set-theoretically from a, c and i in some fixed but arbitrary way, and likewise for $\langle Bel, a, p; i\rangle$, $\langle Tr, p; i\rangle$, $\{s; T\}$ and $[\sigma]$. In saying SOA, SIT, etc. are the greatest classes that satisfy the indicated conditions, we mean that if SOA', SIT', ... also satisfy those conditions, then $SOA' \subseteq SOA$, $SIT' \subseteq SIT$, etc. The existence of SOA, SIT, etc. is proved analogously to the proof that every monotone operator has a greatest fixed point. In fact, if we define $(X_1 \ldots X_n)$, where $X_1 \ldots X_n$ are classes, to be the class $\{(x, i) : i = 1 \ldots n \text{ and } x \in X_i\}$, then the class $(SOA, SIT, PROP, AtTYPE)$ is the greatest fixed point of a suitable monotone operator. Like its Russellian counterpart, $PROP$ contains many self-referential propositions.

If a proposition p is $\{s; T\}$ for some T, then s is said to be the situation that p is *about*. The states of affairs $\langle H, a, c; 0\rangle$ and $\langle H, a, c; 1\rangle$ are said to be *duals* of each other, as are $\langle Bel, a, p; 0\rangle$ and $\langle Bel, a, p; 1\rangle$, and $\langle Tr, p; 0\rangle$ and $\langle Tr, p; 1\rangle$; if σ is a state of affairs, $\overline{\sigma}$ is its dual. The negation $\overline{T}$ of a type T is defined inductively:

- $\overline{[\sigma]} = [\overline{\sigma}]$;
- $[\overline{\bigwedge X}] = [\bigvee\{\overline{p} : p \in X\}]$;
- $[\overline{\bigvee X}] = [\bigwedge\{\overline{p} : p \in X\}]$.

Truth for Austinian propositions is also defined inductively:

- $\{s; [\sigma]\}$ is true iff $\sigma \in s$;
- $\{s; [\bigwedge X]\}$ is true iff $\{s; T\}$ is true for each $T \in X$;
- $\{s; [\bigvee X]\}$ is true iff $\{s; T\}$ is true for at least one $T \in X$.

A proposition is *false* just in case it is not true. We say that s is *of type* T just in case $\{s; T\}$ is true.

An oddity of this formalism is that the truth value of an Austinian proposition depends only on its set-theoretic structure, and not, for example, on which states

of affairs correspond to facts in the world. Barwise and Etchemendy offer two ways around this apparent difficulty. One is to say that real situations, as opposed to their set-theoretic counterparts, would have had different states of affairs as constituents if the world had been different; the apparent independence of a proposition's truth value from how the world is is just an artifact of the modeling. The other is to accept this world-independence and maintain that the role of the world is not to determine what propositions are true, but rather which ones are about real situations. I will have more to say about this later.

In any case, in terms of the mathematical modeling, the world plays no role in determining which propositions are true, but does play a role in determining which ones are about actual situations. To make this precise, a *model of the world* (or simply a *model*) is defined to be a class $\mathfrak{M}$ of states of affairs that satisfies the following coherence conditions:

- No state of affairs and its dual are both in $\mathfrak{M}$;
- $\langle Tr, p; 1\rangle \in \mathfrak{M}$ only if p is true;
- $\langle Tr, p; 0\rangle \in \mathfrak{M}$ only if p is false.

A situation s is said to be *actual* with respect to $\mathfrak{M}$ if $s \subseteq \mathfrak{M}$, and *possible* if it is actual with respect to some model. Equivalently, s is possible just in case s is itself a model. A proposition is *accessible* if it is about an actual situation. The propositions expressed by natural language utterances are all about collections of facts, i.e., states of affairs that actually obtain; so if a model $\mathfrak{M}$ is used to represent the totality of facts, then the expressible propositions should be represented by the propositions accessible w.r.t. $\mathfrak{M}$.

A model is called *total* if it is not properly contained in any model.

THEOREM 2. (1) *A model* $\mathfrak{M}$ *is total iff every state of affairs or its dual belongs to* $\mathfrak{M}$; (2) *if* $\mathfrak{M}$ *is total, then for all propositions* p, $\langle Tr, p; 1\rangle \in \mathfrak{M}$ *iff* p *is true, and* $\langle Tr, p; 0\rangle \in \mathfrak{M}$ *iff* p *is false*; (3) *every model can be extended to a total model.*

PROOF. (2) is immediate from (1) and the definition of model, and the right-to-left direction of (1) is trivial; so suppose $\mathfrak{M}$ is a model and suppose that neither σ nor $\overline{\sigma}$ belongs to $\mathfrak{M}$. If σ is not of the form $\langle Tr, p; i\rangle$, then $\mathfrak{M} \cup \{\sigma\}$ and $\mathfrak{M} \cup \{\overline{\sigma}\}$ are clearly both models. If $\sigma = \langle Tr, p; 1\rangle$, then $\mathfrak{M} \cup \{\sigma\}$ ($\mathfrak{M} \cup \{\overline{\sigma}\}$) is a model if p is true (false), and similarly if $\sigma = \langle Tr, p; 0\rangle$. In any case, then, at least one of $\mathfrak{M} \cup \{\sigma\}$ or $\mathfrak{M} \cup \{\overline{\sigma}\}$ is a model, and therefore $\mathfrak{M}$ is not total.

As for (3), suppose $\mathfrak{M}$ is a model and let $\mathfrak{M}' = \mathfrak{M} \cup \{\sigma : \sigma \text{ is } \langle H, a, c; 1\rangle \text{ or } \langle Bel, a, p; 1\rangle \text{ and } \overline{\sigma} \notin \mathfrak{M}\} \cup \{\langle Tr, p; 1\rangle : p \text{ is true}\} \cup \{\langle Tr, p; 0\rangle : p \text{ is false}\}$; then $\mathfrak{M}'$ is as required. □

REMARK. If $\mathfrak{M}$ contains σ or $\overline{\sigma}$ whenever σ is not of the form $\langle Tr, p; i\rangle$, then $\mathfrak{M}$ has a *unique* total extension.

There is also an Austinian semantics for the language $\mathcal{L}$, which is a fairly straightforward modification of the Russellian semantics. Now, the proposition expressed by a formula depends on the context, which supplies a situation. In particular, if the context specifies a situation s, then in that context the Liar sentence $\neg$**True**(**this**) expresses the proposition $f_s = \{s; [Tr, f_s; 0]\}$.

Now let's look at some Liar propositions and sentences. Throughout we assume a fixed but arbitrary total model $\mathfrak{M}$. A *Liar proposition* is a proposition of the form $p = \{s; [Tr, p; 0]\}$. For every situation s, there is a unique Liar proposition about s, called f_s. As just noted, the Liar sentence $\neg$**True**(**this**) always expresses the Liar proposition f_s when the situation specifies s. The proposition f_s is true iff $\langle Tr, f_s; 0\rangle \in s$. If s is actual, then $\langle Tr, f_s; 0\rangle \in s$ only if f_s is false. Therefore, if s is actual, then f_s is false and the fact of f_s's falsehood, i.e., the state of affairs $\langle Tr, p; 0\rangle$, is not an element of s. Now since the state of affairs $\langle Tr, p; 0\rangle$, despite not belonging to s, does belong to the model $\mathfrak{M}$, there is an actual situation $s' \supseteq s$ that includes $\langle Tr, p; 0\rangle$ (take $s' = s \cup \{\langle Tr, p; 0\rangle\}$, for example). And from this situation s' we may form the proposition $p = \{s'; [Tr, f_s; 0]\}$. Like f_s, p attempts to say that

f_s is false; however, p, unlike f_s, is true, since the fact of f_s's falsehood belongs to the situation that p is about. (Of course, we may also form the Liar proposition $f_{s'} = \{s'; [Tr, f_{s'}; 0]\}$ about s', which is false.)

This is the key to handling the strengthened Liar problem. The sentence

(30) Proposition (30) is not true

is translated into $\mathcal{L}$ as the Liar sentence $\neg$**True**(**this**), which always expresses a false proposition–that is, it expresses a false proposition in any context that specifies an actual situation. (Here I am using 'proposition (n)' as a name for the proposition expressed by sentence (n).) This falsity is recognized in

(31) Proposition (30) is not true

from which we feel we are entitled to conclude

(32) Proposition (31) is true.

This is *prima facie* a contradiction: since (30) and (31) are the same sentence, they intuitively ought to express the same proposition, which would seem to render (31) and (32) mutually inconsistent. Again a shift in context, this time between (30) and (31), explains this away. When (30) is uttered, the context specifies an actual situation s, and (30) expresses the Liar proposition f_s; since s is actual, f_s is false and $\langle Tr, f_s; 0\rangle \notin s$. When (31) is uttered, a new actual situation $s' \supseteq s$ is specified such that $\langle Tr, f_s; 0\rangle \in s'$. Sentence (31), which is translated as (say) $\neg$**True**(**that**$_1$), expresses the true proposition $p = \{s'; [Tr, f_s; 0]\}$, assuming that the context specifies f_s as the referent of '**that**$_1$'. When (32) is uttered, the context specifies an actual situation s'' that contains the fact $\langle Tr, p; 1\rangle$ and specifies the proposition p as the referent of '**that**$_2$', so that (32), which is translated **True**(**that**$_2$), expresses the true proposition $\{s''; [Tr, p; 1]\}$. Thus we see that (30)–(32) may well involve two shifts

in context, namely from s to s' and from s' to s''. It need not involve a second shift, however, for although s' is always distinct from s, s'' and s' may be the very same situation. For example, s' might be the unique situation such that $s' = s \cup \{\langle Tr, f_s; 0\rangle, \langle Tr, \{s'; [Tr, f_s; 0]\}\rangle\}$.

4.4. Problems with *The Liar*. Barwise and Etchemendy's account, as I said, is similar in important respects to Parsons' view, so the objections raised against the latter might well be raised against the former. In fact, I think those objections do pose serious problems for *The Liar*. There is also at least one problem peculiar to *The Liar*, and I will start with that.

Earlier, we saw that an Austinian proposition's truth value seems, curiously, not to depend on how the world is. As I noted, Barwise and Etchemendy suggest that real situations, unlike their mathematical counterparts, have their constituent facts accidentally rather than essentially, i.e., they would have had different constituent states of affairs if the world had been relevantly different. If this is so, then it ought to be possible to modify the formal account so as to capture this world-dependence.

In particular, since models represent the world, it seems natural to arrange things so that a situation determines a set of states of affairs not all by itself, but rather in conjunction with a model. One way to accomplish this is as follows. Define the classes *PreSOA*, *SOA*, *SIT*, *PROP*, and *AtTYPE* to be the largest classes that satisfy the following:

- Every element of *PreSOA* has of one of the following forms:
  - $\langle H, a, c\rangle$
  - $\langle Bel, a, p\rangle$
  - $\langle Tr, p\rangle$

  where $p \in PROP$;
- Every $\sigma \in SOA$ is of the form $\langle x, i\rangle$, where $x \in PreSOA$ and i is 0 or 1;
- Every $s \in SIT$ is a subset of *PreSOA*;

- Every $p \in PROP$ is of the form $\{s; T\}$, where $s \in SIT$ and $T \in \Gamma(AtTYPE)$;
- Every $T \in \Gamma(AtTYPE)$ is $[\sigma]$ for some $\sigma \in SOA$.

Define a *pre-model* to be any class $\mathfrak{M}$ of states of affairs such that $\langle x, 0\rangle$ and $\langle x, 1\rangle$ never both belong to $\mathfrak{M}$. If $s \in SIT$ and $\mathfrak{M}$ is a pre-model, define $s_{\mathfrak{M}}$ to be $\{\langle x, i\rangle : x \in s \text{ and } \langle x, i\rangle \in \mathfrak{M}\}$. Thus, $s_{\mathfrak{M}}$ may be thought of as the set of states of affairs that would constitute the situation s if $\mathfrak{M}$ were actual. Define the relation of truth in a pre-model in the natural way, as follows:

- $\mathfrak{M} \models \{s; [\sigma]\}$ iff $\sigma \in s_{\mathfrak{M}}$;
- $\mathfrak{M} \models \{s; [\bigwedge X]\}$ iff $\mathfrak{M} \models \{s; T\}$ for all $T \in X$;
- $\mathfrak{M} \models \{s; [\bigvee X]\}$ iff $\mathfrak{M} \models \{s; T\}$ for some $T \in X$.

As before, *false in* $\mathfrak{M}$ means not true in $\mathfrak{M}$. We may now define a model to be a pre-model $\mathfrak{M}$ such that whenever $\langle Tr, p; i\rangle \in \mathfrak{M}$, p is true (false) in $\mathfrak{M}$ if $i = 1(0)$.

We may define a model to be *maximal* if it is not properly contained in another model, and *total* if it contains each state of affairs or its dual. This time, however, it turns out that there are no total models. We can see this by letting $s = \{\langle Tr, p\rangle\}$, where $p = \{s; [Tr, p; 0]\}$, and observing that neither $\langle Tr, p; 0\rangle$ nor $\langle Tr, p; 1\rangle$ belongs to any model: if $\langle Tr, p; 0\rangle \in \mathfrak{M}$, then $\langle Tr, p; 0\rangle \in s_{\mathfrak{M}}$ and p is true in $\mathfrak{M}$, contradicting the definition of model; and if $\langle Tr, p; 1\rangle \in \mathfrak{M}$ then $\langle Tr, p; 0\rangle \notin s_{\mathfrak{M}}$ and p is false in $\mathfrak{M}$, again contradicting the definition of model. So on this variant of the Austinian account, every model is partial, in the sense that it is not total. Every model is also partial in the sense that some proposition and its negation are both untrue in it, namely the proposition p and its negation $\{s; [Tr, p; 1]\}$.

Thus we loose much of the simplicity that the unmodified account had as compared to the Russellian account. More importantly, the new account has a harder time handling the strengthened Liar problem. Call a proposition l a *strong Liar* if (for some s) $l = \{s; [Tr, l; 0]\}$ and $\langle Tr, l\rangle \in s$. If l is a strong Liar then it is false in any model, but so is $\{s'; [Tr, l; 0]\}$ for any situation s' whatsoever; thus while a strong Liar is false, there is no true proposition that says it is false, and the contextual shift

that normally vindicates strengthened Liar reasoning is no longer available. That is, let $\mathfrak{M}$ represent the world, and suppose the utterance (30) above expresses a strong Liar. Then (30) expresses a false (in $\mathfrak{M}$) proposition, but now (31) also expresses a false proposition, no matter what contextual shift has occurred between (30) and (31). And no matter what happens to the context between (31) and (32), (32) also says something false. So while the present version of the Austinian account tells us that (30) expresses a false proposition, it gives us no way to correctly conclude that that consequence of the account, that (30) expresses a false proposition, is true.

One possible way of getting around this is to deny that (30) ever expresses a strong Liar proposition. This by itself is not enough, however, because there might still be a singlular term that denotes such a proposition. In that case we are committed to

(33) p is false,

where 'p' is a singlular term denoting a strong Liar, yet it is incorrect to conclude that (33) says something true. And while the formal language $\mathcal{L}$ might lack singular terms that denote strong Liars, it's hard to see how any language in which the account itself can be stated–e.g., English–can lack such terms. Even aside from this point, if the modified Austinian account is right then there are strong Liar propositions and they are all false, yet 'There are strong Liar propositions and they are all false' is presumably incapable of expressing a true proposition given any reasonable way of handling quantification.

Moreover, the difficulties we just encountered arise on a wide range of ways of modifying the original Austinian account. They arise, for example, on any account that satisfies the following:

- Every model is a class of states of affairs;
- If s is a situation and $\mathfrak{M}$ is a model, then $s_{\mathfrak{M}} \subseteq \mathfrak{M}$;
- A proposition $\{s; [\sigma]\}$, where s is a situation and σ is a state of affairs, is true in a model $\mathfrak{M}$ iff $\sigma \in s_{\mathfrak{M}}$;

- If $\mathfrak{M}$ is a model, then $\langle Tr, p; 0\rangle \in \mathfrak{M}$ only if p is false in $\mathfrak{M}$ and $\langle Tr, p; 1\rangle \in \mathfrak{M}$ only if p is true in $\mathfrak{M}$;
- There are a situation s and a proposition p such that $p = \{s; [Tr, p; 0]\}$ and $\langle Tr, p; i\rangle \in s_{\mathfrak{M}}$ iff $\langle Tr, p; i\rangle \in \mathfrak{M}$ for $i = 0, 1$.

If these conditions all obtain, then the proposition p mentioned in the last condition is false in every model, yet $\langle Tr, p; 0\rangle$ does not belong to any model. So the problem of strong Liars is not merely an accident of one particular way of modifying the Austinian account.

Perhaps a better strategy, then, is to place restrictions on which situation-type pairs are to count as propositions, or on which subsets of *PreSOA* are to count as situations, or both. I have no particular objection to doing this, but there are two points that need to be made. First, extending the Austinian account in an acceptable way appears to be a nontrivial task which has not, to my knowledge, been done; the account given in *The Liar* must be regarded as incomplete in a crucial respect. Second, one tends to get the impression on reading *The Liar* that what's doing all the work of handling the paradoxes is the fact that situations are partial. We now see that this is not the case: a pivotal role is also played by the fact that the propositions considered there are, in effect, of a very special type: ones that say that a set s of facts has a certain property T, where whether s has T depends only on what facts do or don't belong to s. Once we broaden our perspective to include propositions whose truth values depend on the way the world is, we must compensate by imposing some new restriction.

Perhaps, then, we should consider Barwise and Etchemendy's other suggestion, that propositions really are true or false independently of how the world is, that the only role the world plays is to determine which propositions are accessible and, hence, to help determine what proposition a given utterance expresses. This is certainly not the way we usually think about propositions, especially when we think of them as what 'true' applies to; I think we would normally say that when something is true,

what makes it true is some feature of the world. In any case, this move won't avoid the problem, since the problem will resurface once we try to give an account of such notions as actuality and accessibility. Suppose, for instance, that we extended the Austinian account to include propositions that say that a given situation is actual. The natural way to proceed would be to add a new kind of state of affairs $\langle Act, s; i\rangle$ to the effect that s is or isn't actual, and to stipulate that if $\mathfrak{M}$ is a model, then $\langle Act, s; 1\rangle \in \mathfrak{M}$ only if s is actual in $\mathfrak{M}$ and $\langle Act, s; 0\rangle \in \mathfrak{M}$ only if s is not actual in $\mathfrak{M}$. As is easily seen, if $\langle Act, s; 0\rangle \in s$ then s is not actual in any model, yet the fact of s's nonactuality, $\langle Act, s; 0\rangle$, does not belong to any model, and the propositions $\{s'; [Act, s; 0]\}$ that say s is nonactual are false for all possible s'. This is precisely the sort of problem we just encountered, and solving it would seem to require some sort of restriction on what sets of states of affairs are to count as situations.

My second reservation about the Austinian account is related to something we already encountered with Parsons. The Liar sentence $\neg$**True**(**this**) always expresses a false proposition p, and the fact of p's falsehood does not belong to the situation p is about. The same goes, presumably, for natural language sentences: 'This proposition is not true' on a given occasion of use expresses a proposition p about a portion of the world that excludes the fact of p's falsehood. But p's falsehood seems to be precisely what this sentence is being used to make a claim about. At any rate, one would think that it *could* be so used, just as 'Claire has the ace of hearts', while it can be used to say something about a particular poker hand, can also be used to say something about Claire, wherever she may happen to be.

In the 1989 postscript to *The Liar*, Barwise and Etchemendy consider essentially this point:

> [I]t may be that the demonstrative conventions presuppose a condition like F-closure [i.e., $\langle Tr, p; 0\rangle \in s$ whenever p is false], a condition that simply cannot be met in the case of an utterance of the Liar sentence. If this is so, then the Austinian account would have a slightly different

> force. It would show that, assuming this demonstrative convention, the Liar sentence could not express a proposition at all: its use would guarantee that the assumed convention is violated. (p. 190)

They regard the matter as part of the pragmatics of natural language and hence beyond the scope of *The Liar*, which is concerned solely with the logical aspects of Liar sentences and propositions. But they may be too quick in consigning this problem to the garbage heap of pragmatics. Handling paradoxical sentences by saying they fail to express propositions is known to cause trouble in other contexts (e.g., gap accounts), via the propositional Liar sentence 'this sentence expresses no true proposition'; so we ought to see whether it also causes the same sort of trouble here.

So let's consider the propositional Liar sentence sentence. Since it is utterances rather than sentence types that express propositions, what we ought to consider is an utterance u of the sentence 'u does not express a true proposition'. Can u express a proposition? It might, if that proposition were about an appropriately restricted situation. But the context of u's utterance could be such that if u expresses any proposition at all, it expresses one about a situation that includes the fact that u expresses a true (false) proposition. In that case u obviously can't express a true proposition; and if u expressed a false proposition p, then since p would be about a situation that includes the fact that p expresses a false proposition, p would be true. Thus u expresses no proposition at all in this case.

By itself, this does not give rise to a strengthened Liar problem. Granted, u does not express a proposition, so *a fortiori* u expresses no true proposition, and granted that u says that u expresses no true proposition. But when we conclude that u expresses no true proposition, we do so by making a new utterance u', and there is no reason why u' should fail to express a true proposition. (If u' expresses a true proposition, it follows that u expresses no true proposition, but it does not follow that u' expresses no true proposition.) Notice, however, that the paradoxes are crucially relevant to whether a given utterance expresses a proposition; contrary

to what Barwise and Etchemendy suggest, the question of whether an utterance expresses a proposition is very appropriate in a study of the paradoxes.

But while the utterance u doesn't pose a problem, there's trouble just around the corner. We run into trouble when we consider utterances of the sentence type

(34) No utterance of (34) expresses a true proposition.

For simplicity, let's imagine that when an utterance u of (34) expresses a proposition p, p is about a situation sufficiently inclusive to include the fact that u expresses p and to include the fact that p is true (or false).[14] Then clearly no utterance of (34) expresses a true proposition. But now we are in real trouble, since expressing this conclusion involves uttering a token of (34), which as we have just seen cannot express a true proposition. The strengthened Liar has not really been avoided. Strengthened Liar problems also arise in a somewhat different way. In the case we considered two paragraphs ago, what allows us to avoid an outright contradiction is the fact that u and u' are different utterances, albeit utterances of the same sentence. However, it is possible to arrange things so that u and u' are one and the same. Imagine that Susan correctly believes that the first sentence token written on the blackboard in room 666 after 12:00 is a token of the following:

> The first sentence token written on the blackboard in room 666 after 12:00 expresses no true proposition.

Call this token t. Suppose she also believes, again correctly, that the intentions behind the writing of t are such as to prevent it from expressing a proposition. She writes this conclusion down by means of a token t' of the same type as t. So far, so good. However, now she learns something new: *t and t' are the very same token.* (Maybe she thought herself to be in room 999; maybe she thought it was 11:00.) Her conclusion, which she wrote down via t', seems to have been justified by the lights of the theory

[14]This simplifying assumption could be eliminated by supposing that we have adopted a convention that allows us to indicate explicitly when we are talking about a situation which is inclusive in this way; (34) could then be rewritten so as to explicitly incorporate our assumption.

now under consideration, yet at the same time that theory tells us that it fails to express a proposition.

Finally, the same statability problem that we encountered with Parsons applies here as well. According to Barwise and Etchemendy, we never speak of the whole world at once, in the sense that we never express a proposition about a situation that includes all the facts. How, then, are we to understand the reference to the 'whole world' in their claim that we never talk about the whole world?

Their answer is completely different from Parsons'. In the main body of *The Liar* they make it quite clear that, according to the Austinian account, we can never make a claim about the whole world, and they do not discuss the issue I just raised. In the postscript to *The Liar*, they take a different line. Formally, propositions are represented by pairs $\{s; T\}$, where s is a set, and total models, which are proper classes, represent the world, so the formal side of the account suggests that a proposition can never be about the whole world. However, Barwise and Etchemendy claim that this is an artifact of the way propositions and the world are modeled, not a consequence of basic ideas of the Austinian account itself. While a Liar proposition cannot be about the whole world, and hence the Liar sentence cannot express a proposition about the whole world, some other propositions may well be about the whole world, and some sentences may express them.

Barwise and Etchemendy say that there are several ways to make formal sense of this idea, and indicate one way. The idea is to find a suitable possible situation s to represent the whole world. (While s will in fact be a set and not a proper class, this fact is irrelevant; s still *represents* the whole world.) Given a situation s, call a proposition p an *s-proposition* if $\langle Tr, p; i\rangle \in s$ for some i, and call a situation s' an *s-situation* if $s' \subseteq s$ and some s-proposition is about s'. When s is taken to represent the world, then the set of s-propositions represents all the propositions there are, and the set of s-situations represents all the situations there are. Since s is a possible situation, the Liar proposition f_s about s is false, but its falsity is not a fact of s; it

follows that f_s is not an s-proposition, and hence, on this way of doing things, is not to be thought of as a proposition at all. There may, however, be other propositions about s that are s-propositions; and in that case s itself will be an s-situation.

Barwise and Etchemendy go on to discuss how an appropriate s might be chosen, using a technical result they call the Reflection Theorem. I won't go into the technical details, though, since there are serious problems with the entire approach. It is hardly believable that while we sometimes speak of the entire world, any utterance of the Liar sentence makes a statement about a proper part of the world. It's one thing to say that we *never* speak of everything; once it is admitted that we do sometimes use language to speak of everything, however, it's hard to resist the conclusion that the Liar could be so used. (One way to see this is to imagine that we have adopted a convention that allows us to indicate explicitly when we are speaking of everything; now consider a Liar sentence in which this explicit indication is made.) The only credible way to avoid Liars with this sort of universality is to maintain that whenever the Liar sentence is used, either it expresses a proposition about less than there is, or it expresses no proposition at all. But the former case is implausible in general, and in the latter case the view will have troube handling propositional Liar sentences, as we saw above.

5. Gaifman

Next, let us briefly examine Gaifman's theory of truth, as put forth in [**Gai88**] and [**Gai92**]. While the technical details of the construction are complicated, the underlying ideas are simple. Gaifman holds truth to be a property of tokens, rather than types.[15] This is the key to handling the strengthened Liar (which Gaifman sharply criticizes Kripke for evading). The token

(35) The sentence on line 35 is not true

is neither true nor false, while the token we use to say that (35) is not true, namely

(36) The sentence on line 35 is not true,

is true, a fact that we could go on to assert by means of a third token

(37) The sentence on line 36 is true.

Thus, as with the other theories discussed in this chapter, we can consistently assertively utter a sentence and say that the sentence is not true, provided the token we utter is different from the one we say is untrue. (Of course, we shouldn't assertively utter a token and claim that that very *token* is untrue.)

Thus, a sentence of the form 'p is true', even when 'p' denotes a token, can be used on some occasions to say something true, and on others to say something untrue, in the sense that some of its tokens are true and some are not true. (More accurately, while tokens like (36) say something true, tokens like (35) fail to say anything at all. So it should not be supposed that the sentence S = 'The sentence on line 35 is true' expresses different propositions on different occasions: there is a single proposition p associated with S, namely S's meaning, and some tokens of S express p while other

[15]More accurately, it is a property of what Gaifman calls *pointers*. Pointers are obtained by enlarging the class of tokens so that, for example, whenever a sentence type φ has a pointer p, its negation $\neg\varphi$ has an associated pointer p'. This is done for the sake of technical convenience, and will not affect anything I have to say in this section.

express no proposition at all.) Notice that the difference between (35) and (36) is not explained in terms of shifting domains of quantification (or situations), or in terms of a truth predicate with a shifting extension. The truth predicate has a single, fixed extension (namely, the set of true tokens), although a token of 't is true' can fail to be true even when t belongs to that set. In fact, there seems to be no hierarchy that plays a role in Gaifman's account analogous to the roles played by hierarchies in other accounts we have been examining; the term 'hierarchy account' is not entirely appropriate here.

It turns out, though, that the kind of objection I have been advancing in this chapter goes through quite smoothly when applied to Gaifman's theory. This is easily seen by considering the sentence type

(38) No token of (38) is true.

Clearly no token of (38) is true, since if (38) had a true token t, it would follow at once that no token of (38) is true, and hence that t is not true. (Here we are using the schema 'If t is a token of 'A' and t is true, then A', which Gaifman accepts.) So no token of (38) is true; but as soon as we say so, we have uttered a token of (38); so Gaifman's theory requires us to assertively utter a sentence token and at the same time to say that that token is not true, which is nothing other than the familiar strengthened Liar problem. So, at least at first blush, Gaifman's account seems to be subject to the strengthened Liar difficulties it seeks to avoid.

It still seems to be subject to them at second blush. None of the strategies we have yet encountered for avoiding this problem apply here. The analogous problem for the Tarskian hierarchy involves the sentence 'This sentence is not true at any level', and the response is that such a sentence attempts to illegitimately bind the subscripts of 'true' by a quantifier, or what comes to the same thing, that the argument involves mistakenly thinking of truth-at-a-level as a relation between a sentence and a level. But no analogous move is available to Gaifman. The sentence (38) is

perfectly well-formed. A Parsons-style move is also unavailable. Parsons' response to the analogous objection is to deny that unrestricted quantification is possible; but the only quantification in (38) is quantification over tokens, and surely there is no problem in quantifying over all of *them*.

To understand what Gaifman would say to all this, we must first understand his notion of a "black hole." First, 'n-hole' is defined inductively: a *0-hole* is a gappy sentence, and an $n+1$-*hole* is an n-hole S such that both 'S is true' and 'S is false' are gappy. A black hole is now defined to be a sentence that is an n-hole for all n. If S is a 1-hole but not a 2-hole, for example, then while information about S cannot be conveyed directly (via sentences like 'S is not true' or 'S is neither true nor false'), it may be conveyed indirectly, e.g., via "S is true' is neither true nor false'. Black holes, however, are semantic untouchables: no semantic information about them can be conveyed, however indirectly. Since we have $\mathfrak{M} \models \varphi$ iff $\mathfrak{M} \models Tr(\text{`}\varphi\text{'})$ and $\mathfrak{M} \dashv \varphi$ iff $\mathfrak{M} \dashv Tr(\text{`}\varphi\text{'})$ for a fixed point partial model $\mathfrak{M}$, every hole in Kripke's construction is a black hole.

The definition of n-hole just given is no longer appropriate if we assume with Gaifman that truth applies to tokens rather than types, since even if p is a token, 'p is true' and 'p is false' are types of which there may be several tokens. The appropriate notion would seem to be this: a 0-hole is a gappy token, and an $n+1$-hole is an n-hole p such that each token of 'p is true' or of 'p is false' is gappy. Thus (35) is a 0-hole but not a 1-hole, and (36) is not even a 0-hole.

In his earlier paper, Gaifman briefly discusses an example similar to (38):

> With unbounded quantification holes can arise provided that the language is sufficiently rich. For example, consider a formalisation of something like "Every pointer which points to me and which says that I am true is not true". This may yield a hole, but not a black hole (if Max asserts of this pointer that it is not true he will get GAP, but then Moritz can assert that Max's assertion is not true and get T). There

> are more sophisticated versions of the evaluation algorithm which prevent holes of this and related forms. Their discussion is postponed to the next paper. [**Gai88**, p. 58]

(The postponed discussion has not yet appeared, to my knowledge.) The example here is not exactly (38), but rather

(39) No token of '(39) is true' is true.

(Notice that (39) is best understood as a token, whereas (38) is a type.) By what is essentially Gaifman's reasoning, we can show that (38), like (39), is not a black hole. To see this, let p be a token of (38), and let q be a token of 'p is not true'. Then p is not true, since if p is true, it follows both that (38) has a true token and that it doesn't. However, nothing prevents q being true. In particular, if q is true, it follows that p is not true, but since q is not a token of (38), it does not follow that (38) has a true token.

This evades the strengthened Liar rather than solving it. What Gaifman has shown is that information about (38) can be conveyed indirectly, via tokens like q. But the strengthened Liar problem isn't about the availability of indirect ways of saying what we cannot say directly. The strengthened Liar, for a given theory of truth, is the problem that the theory has consequences that are untrue according to the theory itself. When truth is ascribed to tokens rather than types, and when our theory has a consequence some, but not all, of whose tokens are true, then perhaps this situation is acceptable. But when a theory has a consequence *all* of whose tokens are untrue (according to the theory), then the theory is self-defeating in precisely the way that simple truth value gap accounts are shown to be self-defeating by the ordinary Liar. And 'no token of (38) is true' is certainly a consequence of Gaifman's theory, yet according to that theory no token of that sentence type is true. In short, questions concerning holes and the ability to convey semantic information indirectly

are simply not relevant to the strengthened Liar, and so Gaifman appears to have no answer to the present version of the strengthened Liar objection.

Gaifman's account is also subject to a difficulty we encountered in the last section. As I noted there, we can construct examples of Liar reasoning in which the initial token, which is claimed to be gappy, and the subsequent token, by which one says the first token is gappy, are the very same token. (Recall the example of Susan in room 666.) More generally, recall from the beginning of this chapter that when presented with a strengthened Liar argument

(a) S is not true (where $S =$ 'S is not true')

(b) 'S is not true' is true

(c) S is true,

the proponent of a hierarchy account has two options. He may claim that the utterances S and (a) express different propositions, in which case (b) should be changed to '(a) is true' and (c) doesn't follow; or he may claim that S and (a) do express the same proposition, and that (c) follows, but that (c) does not contradict (a), since a change in context has occurred. The former may perhaps be more natural. It cannot be the answer in general, though, even putting aside our general worries about the entire hierarchy approach, since the utterances S and (a) simply need not be distinct.

6. Kripke and Soames

Finally, let's look at an account suggested by Kripke's remarks on the limitations of the fixed point approach, and developed in detail by Scott Soames in [**Soa97**]. This account combines the truth value gap account with the hierarchy approach, yielding a hierarchy each of whose levels contains a gappy truth predicate.

6.1. The Account. In [**Kri75**], Kripke writes:

> Now the languages of the present approach contain their own truth predicates and even their own satisfaction predicates, and thus to this extent the hope [for a universal language] has been realized. Nevertheless the present approach certainly does not claim to give a universal language, and I doubt that such a goal can be achieved. First, the induction defining the minimal fixed point is carried out in a set-theoretic metalanguage, not in the object language itself. Second, there are assertions we can make about the object language that we cannot make in the object language. For example, Liar sentences are not true in the object language, in the sense that the inductive process never makes them true; but we are precluded from saying this in the object language by our interpretation of negation and the truth predicate. If we think of the minimal fixed point, say under the Kleene valuation, as giving a model of natural language, then the sense in which we can say, in natural language, that a Liar sentence is not true must be thought of as associated with some later stage in the development of natural language, one in which speakers reflect on the generation process leading to the minimal fixed point. It is not itself a part of that process. The necessity to ascend to a metalanguage may be one of the weaknesses of the present theory. The ghost of the Tarski hierarchy is still with us. (p. 80)

The discussion continues in a footnote:

> [S]uch semantical notions as "grounded", "paradoxical", etc. belong to the metalanguage. This situation seems to me to be intuitively acceptable; in contrast to the notion of truth, none of these notions is to be found in natural language in its pristine purity, before philosophers reflect on its semantics (in particular, the semantic paradoxes). If we give up the goal of a universal language, models of the type presented in this paper are plausible as models of natural language at a stage before we reflect on the generation process associated with the concept of truth, the stage which continues in the daily life of nonphilosophical speakers.

Soames' elaboration of this idea goes basically as follows. English is an open-ended language which we enrich from time to time by adding an expression for a concept not previously expressible. Ordinary English, the language spoken by ordinary speakers of English, contains its own gappy or vague truth predicate, but does not have the resources to say that its own sentences are ungrounded or indeterminate. It is only when philosophers reflect on the paradoxes that they introduce terms to express these notions. The resulting version of English, which we may call "philosophical English", may itself be extended in a similar manner, and this yields a hierarchy of different versions of English. Unlike the other accounts we have examined in this chapter, there is, according to the present account, a *single* language with a *single* truth predicate, without hidden parameters, which is the only language most speakers of English ever use: namely, the lowest language of the hierarchy.

Recall the two different versions of the gap account of truth: the vagueness account, that if S is a paradoxical sentence, both 'S is true' and 'S is not true' must be rejected; and the simple gap account, that paradoxical sentences are neither true nor false, and *a fortiori* are not true. Either approach might be incorporated into the present account. That is, we could say that for paradoxical sentences S of ordinary

English, it would be wrong for an ordinary English speaker to assert either 'S is true' or 'S is not true'. Or, we could maintain that an ordinary English speaker speaks correctly when he says that such an S is not true. In particular, if S is the sentence 'S is not true', then on the latter view, an ordinary English speaker speaks correctly when he says 'S is not true and 'S is not true' is not true'. But such an assertion seems incorrect. (Nor is it plausible to suppose that it is correct in ordinary English but is somehow rendered incorrect in philosophical English by the incorporation of an ungroundedness predicate.) In short, the simple gap account fails to avoid the strengthened Liar, even in the present setting. For this reason, I will follow Soames in assuming the vagueness account.

(The vagueness account also seems to be a closer fit to Kripke's formalism. At least, if we take the fixed points to represent ordinary English, and interpret the true sentences of a fixed point as the sentences a speaker of ordinary English would be correct in asserting, then we must conclude that if S is an ungrounded sentence, neither 'S is true' nor 'S is not true' is correct, since (the translation of) neither sentence is true in such a fixed point.)

Philosophical English has, and ordinary English lacks, the predicate 'determinately true in ordinary English'. It is important to note that it does not have the predicate 'determinately true in philosophical English': that is, philosophical English, like ordinary English, lacks its own determinate truth predicate. We saw in the last chapter that the simple vagueness account (i.e., without the hierarchy) runs into trouble when one considers the determinate Liar sentence $D =$ 'D is not determinately true'. If philosophical English had its own determinate truth predicate, then the present view would be subject to the same criticism as the ordinary vagueness view of truth. For the sake of brevity I will sometimes write the new predicate as 'determinately true' and refer to it as a determinate truth predicate, but it should be understood that 'determinately true' is shorthand for 'determinately true in ordinary English'.

Now presumably philosophical English has its own truth predicate, just like ordinary English. Is this accomplished by the addition of a new truth predicate, in addition to a predicate for determinate truth in ordinary English? Or does the ordinary English word 'true' automatically serve as a truth predicate for the new language when the determinate truth predicate is added? The latter supposition is far more natural. In general, when we enlarge the English lexicon, thereby creating new English sentences, the word 'true' automatically applies to these new sentences. If, for example, we introduce the word 'glub' to mean blue, then we are automatically licensed in asserting that 'the sky is glub' is true. For this reason, I will assume that a new truth predicate is not added, but rather that the original word 'true' now serves as a truth predicate in philosophical English. This is an issue we will return to later.[16]

Finally, the process of enlarging English needn't, and doesn't, stop with philosophical English. Let us call ordinary English 'English$_1$' and philosophical English 'English$_2$'. In reflecting on the semantics of English$_2$, philosophers may introduce a new predicate meaning *determinately true in English*$_2$, thereby producing yet another version of English, English$_3$. Likewise, English$_3$ may be further expanded to English$_4$, and so on. Thus, we see that English is (at least potentially) an infinite hierarchy of languages.

What happens if we try to run the determinate Liar argument we ran against the unmodified vagueness view in Chapter 3? Recall that the argument relied on two things: the theory's claim that the determinate Liar

(D) (D) is not determinately true

is not determinately true, together with the intuitively valid rules of inference

$$\text{RD}_1\text{:}\ \frac{\vdash \varphi}{\vdash D(\text{`}\varphi\text{'})} \qquad \text{RD}_2\text{:}\ \frac{\vdash D(\text{`}\varphi\text{'})}{\vdash \varphi}$$

[16]Here I am departing from Soames, for whom the hierarchy has different truth predicates on each level as well as different determinate truth predicates.

(Recall that as with RT_1-RT_4, these rules apply only to assertions.) With that assumption and these rules, we saw that it is possible to derive a contradiction.

Assuming the hierarchy of languages, however, we have not one determinate truth predicate, but infinitely many. And if 'D' is interpreted as 'determinately true in English$_n$', then obviously the φ in RD_1 and RD_2 will have to be restricted to sentences of English$_n$. Likewise there is not one determinate Liar now, but infinitely many. And if 'determinately true' is understood to mean 'determinately true in English$_n$', then the determinate Liar is a sentence of English$_{n+1}$, and *not* of English$_n$. Thus, the rules RD_1 and RD_2 simply fail to apply to the determinate Liar. This should not be surprising; formally, the method of avoiding paradox here is essentially the same as that of the Tarskian hierarchy or of Burge.

6.2. Problems. Like the other versions of the hierarchy approach, this version has its problems. To begin with, recall that English$_1$ has a predicate meaning *true* (namely, 'true'), but it does not have a predicate meaning *determinately true*. I think this is hard to maintain if the analogy between 'true' and 'bald' is taken seriously. It seems that with every other vague predicate P, the ordinary speaker of English has a way to say of a given object that it is a determinate case of P, or a determinate non-case, or a borderline case. If Jack is borderline bald, for example, the ordinary speaker of English, not just the philosopher, has a way of saying that Jack is borderline bald.

Now 'true' is probably not vague in the fullest sense of the word; when people say that 'true' is vague, what they really mean is that it has one thing in common with vague predicates, namely *partiality*: there are some things such that to say that they are true, or to say that they are not true, would be incorrect. This may seem to rob some of the force of the argument just given, since it relies on an analogy between 'true' on the one hand and 'bald' or 'poor' on the other, an analogy that may not be very tight. I don't think it robs it of very much force, however. If we have a way of saying, in ordinary English, that Ralph is borderline poor, then this is most plausibly viewed as a general way of saying that a given partial predicate P does or doesn't

clearly apply to a given object, applied to the predicate 'poor' and the object Ralph; it doesn't seem to be specific to vague, as opposed to partial, predicates.

Moreover, the word 'determinately' in 'determinately bald' or 'determinately poor' is most naturally understood as a semantically significant component of those expressions, which can combine with other predicates (such as 'true') to form meaningful expressions. If so, and if 'determinately poor', etc. belong to ordinary English, then the conclusion that 'determinately true' also belongs to ordinary English is unavoidable. This point is easy to miss if we concentrate too much on formal analogues to natural languages and too little on the natural languages themselves. In formal discussions of the paradoxes, natural language is typically represented by means of a formalized language not very different from those we learned about in beginning logic classes. These languages don't have anything closely analogous to adverbs; 'ill' and 'violently ill', for example, are represented by completely different formalized predicates. Thus 'true' and 'determinately true' also wind up getting represented by completely different predicates. Once this happens, it's easy to think that 'true' and 'determinately true' belong to different fragments of English spoken by different populations. But if I am right about the word 'determinately', this is no more plausible than that there is a population of English speakers whose language contains 'ill' and 'violently', but not 'violently ill'.

Another bad reason to think the notion of determinate truth is not to be found in ordinary language comes from confusing technical semantic notions with their informal counterparts. In the technical sense of the word, a grounded sentence (relative to a model $\mathfrak{M}$ for a language $\mathcal{L}$) is a sentence of the language $\mathcal{L} \cup \{T\}$ that belongs to the extension or antiextension of T in the least fixed point over $\mathfrak{M}$. The notion of determinate truth is an informal notion which the technical notion of groundedness may be used to model. They are by no means the same notion: to say that a natural language sentence is determinately true simply does not mean that it is grounded. So the claim that determinate truth is a notion invented by philosophers may seem more

plausible if we conflate it with the technical notion of groundedness than it otherwise would, or than it should.

There are also problems more closely tied to the paradoxes themselves. To begin with, there is the usual statability objection: how the vagueness-hierarchy theory could be statable if true is no clearer than how any of the other theories we have examined could be. As usual, there is nothing incoherent in saying that the account is true of some language; what's incoherent is to purport to say in English that the account is true of English.

Let's verify this in a bit more detail. A formulation of the hierarchy theory would look something like this:

> (*) English consists of an infinite collection of languages, which we will call English$_1$, English$_2$, ..., each of which is a stage in the development of English. For each n, the word 'true' is a (partially defined) truth predicate for English$_n$, and English$_{n+1}$ has (but English$_n$ lacks) a determinate-truth-in-English$_n$ predicate; indeed, English$_{n+1}$ is the language obtained from English$_n$ by adding such a predicate.

(*) certainly seems to be stated in English. If it is, and if the hierarchy account is correct, then it must belong to English$_k$ for some k. In that case, 'is a truth predicate for English$_n$' and 'is a determinate-truth-in-English$_n$ predicate', where in each case n occurs as a variable, are expressions of English$_k$. What could it be to say that a predicate P of English$_n$ is a truth predicate for English$_n$? Presumably it is to say that P applies to all and only the true sentences of English$_n$. But 'true sentence of English$_n$', for variable n, is not expressible in any of the languages of the hierarchy, on pain of paradox.

This argument might be blocked if we found a statement of the theory other than (*), or if we had a way of making sense of that passage without using the general notion of a true sentence of English$_n$. As an example of the second strategy, one could claim that to say English$_n$ has its own truth predicate is simply to say

that there is an expression of English$_n$ that means *true*. (I am assuming here that the meaning of 'true' doesn't change as we ascend the hierarchy, even though its determinate extension and antiextension do.) However, if we try to use this idea to explain what it is for English$_{n+1}$ to have a determinate-truth-in-English$_n$ predicate, we will be in trouble. This is because the statement that P is a determinate-truth-in-English$_n$ predicate iff P means *determinately true in English$_n$* makes no sense unless 'determinately true in English$_n$' makes sense, i.e., the statement can only be made in a language that has a determinate-truth-in-English$_n$ predicate to begin with.

In short, there seems to be no advantage here over, say, Burge's account regarding the statability issue. There is also one more problem, with a rather different flavor. Namely, I will argue, the determinate Liar problem arises even in the lowest levels of the hierarchy.

The main thing to notice is that in general, the truth predicate of ordinary English applies to propositions, not sentences, and that it applies to propositions not expressible in English as well as those that are. The latter point should be obvious, but to see that it holds, observe that if every proposition God believes is true, it does not follow that every proposition God believes is expressible in English.[17] In particular, in contrast to what was said above, the extension of 'true' does not change as we move from English$_1$ to English$_2$: if we are willing to count a sentence true when we speak English$_2$, we should regard it has having already belonged to the extension of 'true' in English$_1$, even if it is not itself a sentence of English$_1$.

Now the hierarchy view depends essentially on the assumption that a given level of English lacks its own determinate truth predicate even though it has its own truth predicate. In particular, English$_1$ lacks its own determinate truth predicate according to that theory, while English$_2$ has a determinate truth predicate for English$_1$ (but not for itself). Yet, using the observations made in the last paragraph, together with

[17]For my purposes, it comes to the same thing to say that 'true' applies to sentences of arbitrary languages, rather than just English.

some plausible assumptions, we can find, in English$_1$, a determinate truth predicate for English$_1$, assuming one exists in English$_2$; this contradicts the basic assumptions of the hierarchy view.

To see how this works, let S be any sentence of English$_1$. Then there is a sentence S^* of English$_2$ that says S is determinately true in English$_1$. Now while S^* is not translatable into English$_1$, English$_1$ may well contain a definite description D that picks out S^*. But since English$_1$'s truth predicate applies to sentences of any language, we can get the effect, in English$_1$, of asserting S^*, namely by saying 'the D is true'.[18]

More generally, let f be a function that takes an English$_1$ sentence S to an English$_2$ sentence $f(S) = S^*$ that says S is determinately true. If f is expressible in English$_1$, via some predicate $F(x, y)$, say, then the expression 'the y such that $F(x, y)$ is true' is effectively a determinate truth predicate for English$_1$. So the hierarchy theory will come to grief if the predicate $F(x, y)$ exists.

And there is every reason to think $F(x, y)$ exists. The process by which the expression 'determinately true in English$_1$' is introduced into English is fully describable in English$_1$, so the proposition that S is determinately true may be picked out in English$_1$ by a description like 'the proposition expressed in English by 'S is determinately true in English$_1$' once 'determinately true in English$_1$' is added to English via such-and-such process'. To make this concrete, consider the way 'determinately true' is added to English: an author simply starts talking about how some sentences are "correct" or "incorrect", how some sentences are such that we "must reject" both them and their negations, etc., and the reader gets the idea readily enough. (Indeed, the reader picks up on the notion of determinate truth so readily that one wonders whether the author has indeed introduced a genuinely new concept, or simply set up

[18]In terms of propositions, S^* expresses a proposition p that is not expressible in English$_1$; but if English$_1$ has a definite description D' that picks out p, we can get the effect of asserting p in English$_1$ by saying 'the D' is true'.

a context which allows a concept that was already expressible in ordinary English to be expressed unambiguously.)

Another way to pick out in ordinary English the proposition that S is determinately true is via actual, flesh-and-blood speakers of philosophical English. If Bob speaks philosophical English, then the proposition that S is determinately true is denoted by the definite description 'the proposition expressed in Bob's ideolect by 'S is determinately true in English_1''. In general, in order to pick out, in English_1, the proposition expressed by an arbitrary sentence of English_2, we really only need to be able to refer in English_1 to the language English_2 itself, for then that proposition can be described, in English_1, as the proposition expressed in English_2 by the given sentence. Again, one interpretation of this fact is that English_2 is in fact not expressively richer than English_1; whether or not this conclusion is correct, the hierarchy of languages does not provide a way out of the determinate Liar problem.

It may be instructive to see in detail how the determinate Liar problem arises in this setting. It will be useful to assume that the function f takes S to a sentence $S^\dagger$ of English_2 that says that S is not determinately true. Thus, suppose 'the F' is a definite description of English_1 that denotes the English_2 sentence

(40) 'The F is true' is not determinately true in English_1.

As with the original definite Liar, it can be shown that, in the presence of the expected disquotational rules, both the assumption that 'The F is true' is determinately true in English_1 and the assumption that it isn't lead to contradictions. (The reasoning that leads to this contradiction takes place in English_2.) These rules are

(RD) $$\frac{A}{\text{`}A\text{' is determinately true in English}_1} \quad \text{and conversely}$$

where A is a sentence of English_1, and

(RT) $$\frac{B}{\text{`}B\text{' is true}} \quad \text{and conversely}$$

where B is a sentence of English$_2$.

The rule (RD) deserves some comment. An (RD) inference takes place in the (allegedly) richer language English$_2$, yet from the assertion A of English$_2$ we conclude that A is determinately true in the language English$_1$. We may do this because English$_2$ was formed by simply adding a new expression, so if a sentence is composed solely from old expressions, it should be true in the new language just in case it is true in the old one. In particular, since the extension of 'true' doesn't change when we move from English$_1$ to English$_2$, (RD) holds when 'A' is the sentence 'the F is true'.

Recall that 'the F' refers in English$_1$ (and hence in English$_2$) to the English$_2$ sentence (D) = ''The F is true' is not determinately true in English$_1$'. From an assertion that 'the F is true' is determinately true in English$_1$, we may reason as follows:

(1) 'The F is true' is determinately true in English$_1$
(2) The F is true (from 1 by (RD))
(3) 'S is not determinately true in English$_1$' is true
(4) S is determinately true in English$_1$ (from 3 by (RT))

i.e., we may derive a contradiction. Similarly, from an assertion that S is not determinately true in English$_1$ we may reason as follows:

(1) 'The F is true' is not determinately true in English$_1$
(2) ''The F is true' is not determinately true in English$_1$' is true (from 1 by (RT))
(3) The F is true
(4) 'The F is true' is determinately true in English$_1$ (from 3 by (RD))

and again we have a contradiction. So we must not assert that the sentence S = 'the F is true' is determinately true (in English$_1$), nor may we assert that it is not determinately true. And as before, we also must not assert that 'S is determinately

true' is itself indeterminate (though showing this formally requires a few assumptions beyond (RT) and (RD)).

One might doubt that '(not) determinately true in English_1' has really been shown to be expressible in English_1, since the English_1 expression 'the y such that $F(S, y)$ is true' arguably does not *mean* that S is (not) determinately true. The formal treatment just given shows that this doesn't matter: we simply don't have to take a stand on this issue.

Finally, it should be mentioned that the only thing this argument relies on (aside from the disquotational rules) is the assumption that determinately true sentences of English_2 already belong to 'true''s extension in English_1; this assumption, in turn, was supported by the fact that, in ordinary English, the word 'true' applies to sentences of all other languages (or equivalently to propositions not expressible in ordinary English), together with the hierarchy theory's identification of ordinary English with the language English_1. Thus the foregoing argument does not work against the other hierarchy accounts examined in this chapter, since those do not identify ordinary English with the bottom level of their hierarchy. However, the foregoing *is* effective against a certain variant of those accounts. A possible response to the statability objection I have been developing against those accounts is that, in stating a given hierarchy account, we are not speaking ordinary English, but rather an essentially richer language spoken only by those who have reflected on the paradoxes. If the foregoing argument is correct, this escape route is not available.

Appendix: Burge's Construction

Burge offers four "constructions," C1–C3 and a modification of C3. In section 2, I made some vague remarks about how one of these (namely, C3) can be regarded as picking out a certain iterated version of the Kripke fixed-point construction. Here I will state and prove the relation between Burge and Kripke more precisely, after which I will examine the modified version of C3. (C1 and C2 are really just simple modifications of the Tarskian hierarchy of languages, so I will not deal with them here.) WARNING: the reader may want to skip the proofs on first reading this appendix.

Each construction is intended to pick out a formal language, although what Burge gives us looks more like a formal deductive system, or rather a list of enunciations about truth that are written in a more or less formalized language. I will take the following approach: I assume as given a classical model $\mathfrak{M}$ for a language $\mathcal{L}$, which we may think of as expressing the nonsemantic notions; we will form a language $\mathcal{L}^* \supseteq \mathcal{L}$, containing the semantic vocabulary that Burge seeks to elucidate and in which the axioms of the various constructions can be written. We assume that $\mathfrak{M}$'s domain includes all the formulas, terms, etc. of $\mathcal{L}^*$, and that $\mathfrak{M}$ can express all the syntactic notions that we will need.[19] In Section 1 we will examine the various expansions of $\mathfrak{M}$ to $\mathcal{L}^*$ that satisfy axioms of C3; that such expansions exists constitutes a proof of C3's consistency, which is something Burge doesn't offer. We will also determine the class of models that C3 picks out, and it will turn out to be probably not exactly what Burge had in mind; we will then see how to fix up C3 so that it picks out the intended class. In section 2 we will give C4 a similar treatment.

A.1. Construction C3. Burge's basic semantic notions are of a formula (of $\mathcal{L}^*$) being *satisfied* by a variable assignment at a level, and of a formula being *rooted* at a level relative to a variable assignment. Intuitively, a formula is rooted at a

[19] I will take a rather carefree attitude toward $\mathfrak{M}$, generally assuming that it has whatever resources are needed for the matter at hand. This is perhaps forgiveable in that our goal is not to establish results about the broadest possible range of mathematical structures, but to describe a certain formal treatment of natural language.

level i if it is grounded relative to the nonsemantic facts together with the facts about satisfaction and rootedness at levels lower than i. For the sake of perspicuity, I have reformulated Burge's constructions in terms of *truth* at a level and the rootedness of a *sentence* at a level. Formally, $\mathcal{L}^* = \mathcal{L} \cup \{T_1, T_2, \ldots\} \cup \{R_1, R_2, \ldots\}$, where the Ts and Rs are distinct unary predicates not in $\mathcal{L}$.

Having said all this, we can now present the axioms of C3, reformulated as I have indicated:

(0) $R_i(\text{`}\varphi\text{'})$, where all the subscripts in φ are less than i.

(1) If $R_i(\text{`}\varphi\text{'})$ then $R_i(\text{`}T_i(\text{`}\varphi\text{'})\text{'})$ and $R_i(\text{`}R_i(\text{`}\varphi\text{'})\text{'})$.

(2) If $R_i(\text{`}\varphi\text{'})$ then $R_i(\text{`}\neg\varphi\text{'})$.

(3) If $R_i(\text{`}\varphi\text{'})$ and $R_i(\text{`}\psi\text{'})$ then $R_i(\text{`}\varphi \to \psi\text{'})$; if $T_i(\text{`}\neg\varphi\text{'})$ or $T_i(\text{`}\psi\text{'})$ then $R_i(\text{`}\varphi \to \psi\text{'})$.

(4) If $R_i(\text{`}\varphi(\overline{x})\text{'})$ for all $x$, then $R_i(\text{`}\forall x\, \varphi\text{'})$; if $T_i(\text{`}\neg\varphi(\overline{x})\text{'})$ for some $x$, then $R_i(\text{`}\forall x\, \varphi\text{'})$.

(5) $R_i(\text{`}\varphi\text{'})$ holds iff it holds by (0)–(4).

(6) $T_i(\text{`}\varphi\text{'}) \to R_i(\text{`}\varphi\text{'})$.

(7) $R_i(\text{`}\neg\varphi\text{'}) \to \big[T_i(\text{`}\neg\varphi\text{'}) \leftrightarrow \neg T_i(\text{`}\varphi\text{'})\big]$.

(8) $R_i(\text{`}\varphi \to \psi\text{'}) \to \big[T_i(\text{`}\varphi \to \psi\text{'}) \leftrightarrow (T_i(\text{`}\varphi\text{'}) \to T_i(\text{`}\psi\text{'}))\big]$.

(9) $R_i(\text{`}\forall x\, \varphi\text{'}) \to \big[T_i(\text{`}\forall x\, \varphi\text{'}) \leftrightarrow \forall x\, T_i(\text{`}\varphi(\overline{x})\text{'})\big]$.

(10) $R_i(\text{`}T_j(\text{`}\varphi\text{'})\text{'}) \to \big[T_i(\text{`}T_j(\text{`}\varphi\text{'})\text{'}) \leftrightarrow T_j(\text{`}\varphi\text{'})\big]$.

(11) $T_i(\text{`}\varphi\text{'}) \to T_k(\text{`}\varphi\text{'}), \quad k > i$.

(12) $T_i(\text{`}\varphi\text{'}) \to T_k(\text{`}T_i(\text{`}\varphi\text{'})\text{'}), \quad k \geq i$.

Notice that (12) follows from the other axioms.

Before proceeding any further, we should note certain omissions from this list. For instance, we certainly want for each closed term t and each sentence φ

$$t = \text{`}\varphi\text{'} \to \big[T_i(\text{`}T_j(t)\text{'}) \leftrightarrow T_i(\text{`}T_j(\text{`}\varphi\text{'})\text{'})\big] \tag{13}$$

and

$$t = \text{`}\varphi\text{'} \rightarrow \left[T_i(\text{`}R_j(t)\text{'}) \leftrightarrow T_i(\text{`}R_j(\text{`}\varphi\text{'})\text{'})\right] \tag{14}$$

which do not follow from (0)-(12). For the same reason, we need to rewrite (1) as

$$R_i(t) \rightarrow \left[R_i(\text{`}T_i(t)\text{'}) \wedge R_i(\text{`}R_i(t)\text{'})\right] \tag{1}$$

and (10) should be rewritten to include atomic sentences not of the form $T_j(\text{`}\varphi\text{'})$:

$$R_i(\text{`}\varphi\text{'}) \rightarrow \big(T_i(\text{`}\varphi\text{'}) \leftrightarrow \varphi\big), \quad \varphi \text{ any atomic sentence.} \tag{10}$$

Finally, only a sentence can be true, and it makes sense to require this explicitly:

$$\forall x \left[T_i(x) \rightarrow \mathit{Sent}(x)\right]. \tag{15}$$

All of these sentence schemas are straightforwardly first order, except for (5). The most natural way to render (5) formally is as the second order schema

$$\forall x \left[R_i(x) \rightarrow \forall X \left(\Phi_i(X) \rightarrow X(x)\right)\right]$$

where $\Phi_i(X)$ is the result of conjoining the ith instances of (0)-(4) and replacing the occurrences of R_i outside quotes with X. Then (5) holds in an expansion $\mathfrak{M}'$ of $\mathfrak{M}$ to $\mathcal{L}^*$ just in case the set $\{\varphi : \mathfrak{M}' \models R_i(\text{`}\varphi\text{'})\}$ is the least set that satisfies $\Phi_i(X)$ in $\mathfrak{M}'$, for each i. (It's easy to show that there always is a least set satisfying $\Phi_i(X)$.) There doesn't seem to be any reason to think that (5) could be formalized as a first order scheme. Since C3 is a theory of truth only for the first order sentences of $\mathcal{L}^*$, it cannot be regarded as a theory of truth for the language in which it is stated.[20]

[20] A certain degree of caution is also needed in interpreting the other sentences of C3. The subscripts must be understood as schematic letters rather than variables, for obvious reasons. However, just about everything else must be interpreted as a variable, since for each i, $\Phi_i(X)$ must be a single *sentence* that says X satisfies (0)–(4) (otherwise the intended meaning of (5) is lost). Thus (7), for example, should be understood as an abbreviation of (some formalization of) 'If the negation of a sentence x is rooted$_i$, then x's negation is true$_i$ just in case x is not true$_i$.'

This also means that the 't' of sentences (1), (13) and (14) needs to be understood as an object-language variable ranging over terms. This means that to formalize (13) and (14) we need to have,

Let's call a classical expansion of $\mathfrak{M}$ to $\mathcal{L}^*$ a *Burge expansion* of $\mathfrak{M}$ to $\mathcal{L}^*$, or a *Burge model*, if it satisfies all the axioms of (reformulated) C3. We will show that there are Burge models and relate them to Kripke's fixed-point construction. First we need a couple of lemmas.

LEMMA 1. *If $\mathfrak{M}'$ is a Burge expansion of $\mathfrak{M}$, then the following hold in $\mathfrak{M}'$ for all φ, ψ and t in $\mathcal{L}^*$ and all i*:

$$\begin{aligned}
R_i(\text{`}\neg\varphi\text{'}) &\leftrightarrow R_i(\text{`}\varphi\text{'});\\
R_i(\text{`}\varphi \rightarrow \psi\text{'}) &\leftrightarrow \big[\big(R_i(\text{`}\varphi\text{'}) \wedge R_i(\text{`}\psi\text{'})\big) \vee T_i(\text{`}\neg\varphi\text{'}) \vee T_i(\text{`}\psi\text{'})\big];\\
R_i(\text{`}\forall x\, \varphi\text{'}) &\leftrightarrow \big[\forall x\, R_i(\text{`}\varphi(\overline{x})\text{'}) \vee \exists x\, T_i(\text{`}\neg\varphi(\overline{x})\text{'})\big];\\
R_i(\text{`}T_i(t)\text{'}) &\leftrightarrow R_i(\text{`}R_i(t)\text{'})\\
&\leftrightarrow R_i(t)
\end{aligned}$$

and the following holds for all i, j with $i < j$:

$$\neg R_i(\text{`}T_j(t)\text{'}) \wedge \neg R_i(\text{`}R_j(t)\text{'})$$

PROOF. If $\mathfrak{M}'$ is a Burge expansion of $\mathfrak{M}$, then R_i's extension in $\mathfrak{M}'$ is the least set satisfying $\Phi_i(X)$ in $\mathfrak{M}'$, and the sets satisfying $\Phi_i(X)$ in $\mathfrak{M}'$ are simply the closed

in the object language, some way of talking about the denotation of an arbitrary term. Fortunately, there is no problem assuming that denotation is expressible in $\mathfrak{M}$—in particular, there is no semantic paradox here. So let us simply assume that denotation is expressible in $\mathfrak{M}$.

points of a certain monotone operator Γ, namely:

$$\begin{aligned}\Gamma(S) = S \cup \{&\text{sentences only containing subscripts } < i\}\\ &\cup \{T_i(\text{`}\varphi\text{'}) : \varphi \in S\}\\ &\cup \{R_i(\text{`}\varphi\text{'}) : \varphi \in S\}\\ &\cup \{\neg\varphi : \varphi \in S\}\\ &\cup \{\varphi \to \psi : \text{ either } \varphi, \psi \in S \text{ or } \mathfrak{M}' \models T_i(\text{`}\varphi\text{'}) \text{ or } \mathfrak{M}' \models \neg T_i(\text{`}\varphi\text{'})\}\\ &\cup \{\forall x\, \varphi(x) : \text{ either } \varphi(\overline{x}) \in S \text{ for all } x \text{ or } \mathfrak{M}' \models \neg T_i(\text{`}\varphi(\overline{x})\text{'}) \text{ for some } x\}.\end{aligned}$$

The least closed point of a monotone operator is also a fixed point, and any fixed point of of Γ evidently satisfies the first four sentences of the lemma. The least fixed point of Γ clearly satisfies the last sentence as well. □

Let C3$^-$ be C3 minus axiom schema (5).

LEMMA 2. *If $\mathfrak{M}'$ is an expansion of $\mathfrak{M}$ to $\mathcal{L}^*$ that satisfies* C3$^-$ *as well as the sentences mentioned in Lemma* 1, *then*

(a) *For all φ and i the sentence $R_i(\text{`}\varphi\text{'}) \to [T_i(\text{`}\varphi\text{'}) \leftrightarrow \varphi]$ holds in $\mathfrak{M}'$*

(b) *For all φ and i the sentence $R_i(\text{`}\varphi\text{'}) \leftrightarrow [T_i(\text{`}\varphi\text{'}) \vee T_i(\text{`}\neg\varphi\text{'})]$ holds in $\mathfrak{M}'$.*

PROOF. (a) follows from (6)-(10) and the sentences of Lemma 1 by a simple induction on the complexity of sentences.

As for (b): if $R_i(\text{`}\varphi\text{'})$ holds in $\mathfrak{M}'$ then so does $R_i(\text{`}\neg\varphi\text{'})$, and we have both $T_i(\text{`}\varphi\text{'}) \leftrightarrow \varphi$ and $T_i(\text{`}\neg\varphi\text{'}) \leftrightarrow \neg\varphi$ true in $\mathfrak{M}'$ by (a); so $\mathfrak{M}' \models R_i(\text{`}\varphi\text{'}) \to T_i(\text{`}\varphi\text{'}) \vee T_i(\text{`}\neg\varphi\text{'})$. Conversely, If $T_i(\text{`}\varphi\text{'})$ holds in $\mathfrak{M}'$ then $\mathfrak{M}' \models R_i(\text{`}\varphi\text{'})$ by (6); and if $\mathfrak{M}' \models T_i(\text{`}\neg\varphi\text{'})$ then $\mathfrak{M}' \models R_i(\text{`}\neg\varphi\text{'})$ by (6), and so $\mathfrak{M}' \models R_i(\text{`}\varphi\text{'})$ as $\mathfrak{M}'$ satisfies the sentences of Lemma 1. So $\mathfrak{M}' \models T_i(\text{`}\varphi\text{'}) \vee T_i(\text{`}\neg\varphi\text{'}) \to R_i(\text{`}\varphi\text{'})$. □

In view of Lemma 2(b), we will treat $R_i(x)$ as an abbreviation for $T_i(x) \vee T_i(\dot{\neg}\, x)$[21] rather than a primitive predicate from now on. Let $\mathcal{L}^+ = \mathcal{L} \cup \{T_1, T_2, \ldots\}$. Reformulating (0)-(15) in $\mathcal{L}^+$ and defining 'Burge expansion of $\mathfrak{M}$ to $\mathcal{L}^+$' are straightforward. We may elimitate occurrences of R_i inside quotes altogether, e.g., (1) is now $R_i(t) \rightarrow R_i(\text{`}T_i(t)\text{'})$, and (14) may be eliminated entirely. $\Phi_i(X)$ is now the result of conjoining the new versions of (0)-(4) and replacing each occurrence outside quotes of $T_i(x) \vee T_i(\dot{\neg}\, x)$ by $X(x)$, and (5) is written in terms of $\Phi_i(X)$ just as before. Lemmas 1 and 2 carry over easily.

Let us now relate Burge models to Kripke's construction. For each $n \geq 0$, let $\mathcal{L}_n = \mathcal{L} \cup \{T_1 \ldots T_n\}$. Let $\mathfrak{M}^+$ be any classical expansion of $\mathfrak{M}$ to $\mathcal{L}^+$, and let $\mathfrak{M}_n$ be $\mathfrak{M}^+$'s $\mathcal{L}_n$-reduct. If $\mathfrak{N}$ is a partial or classical model for $\mathcal{L}_n$ and if E is disjoint from A, let $(\mathfrak{N}, (E, A))$ be the expansion of $\mathfrak{N}$ to $\mathcal{L}_{n+1}$ that assigns E and A to T_{n+1} as extension and antiextension, and let $(\mathfrak{N}, E) = (\mathfrak{N}, (E, -E))$. ($-E$ is E's complement in $\mathfrak{N}$'s domain.) Let σ be a monotone valuation scheme. If $\mathfrak{M}'$ is a partial model for $\mathcal{L}_n$, let us say that $\mathfrak{M}'$ is a *T_n-fixed point* for σ if the extension of T_n in $\mathfrak{M}'$ is $\{\varphi : \mathfrak{M}' \models_\sigma \varphi\}$ and its antiextension in $\mathfrak{M}'$ is $\{\text{nonsentences}\} \cup \{\varphi : \mathfrak{M}' \nvDash_\sigma \varphi\}$.[22] Call $\mathfrak{M}^+$ a *σ-superstructure* over $\mathfrak{M}$ if for each $n > 0$ there is a T_n-fixed point for σ of the form $(\mathfrak{M}_{n-1}, (E, A))$ such that T_n's extension in $\mathfrak{M}^+$ is E. Define $E^{\#} = \{\text{nonsentences}\} \cup \{\varphi : \neg\varphi \in E\}$; then $\mathfrak{M}^+$ is a σ-superstructure just in case $(\mathfrak{M}_{n-1}, (E_n, E_n^{\#}))$ is a T_n-fixed point for each n, where E_n is T_n's extension in $\mathfrak{M}^+$. If $(\mathfrak{M}_{n-1}, (E_n, E_n^{\#}))$ is always the least such fixed point, we will say that $\mathfrak{M}^+$ is the *grounded σ-superstructure* over $\mathfrak{M}$. It is clear that the grounded superstructure over $\mathfrak{M}$ always exists and is unique. In what follows we always use the strong Kleene scheme.

THEOREM 3. *The expansions of $\mathfrak{M}$ to $\mathcal{L}^+$ satisfying the antecedent of Lemma 2 are precisely the superstructures over $\mathfrak{M}$.*

[21] $\dot{\neg}$ is a function symbol of $\mathcal{L}$ that stands for the operation of negation.
[22] By 'nonsentence' we mean anything not a sentence of $\mathcal{L}_n$.

PROOF. It is easy to verify that every superstructure over $\mathfrak{M}$ satisfies the antecedent of Lemma 2. The only not-entirely-trivial case is axiom (11). To see that (11) holds in any superstructure, let $\mathfrak{M}^+$ be a superstructure and suppose $\mathfrak{M}^+ \models T_i(\text{`}\varphi\text{'})$ and $k > i$. Then $\mathfrak{M}_i \models T_i(\text{`}\varphi\text{'})$, $\mathfrak{M}_i = (\mathfrak{M}_{i-1}, E_i)$ and $\mathfrak{N}_i = (\mathfrak{M}_{i-1}, (E_i, E_i^{\#}))$ is a T_i-fixed point. $\mathfrak{N}_i \models T_i(\text{`}\varphi\text{'})$, and hence $\mathfrak{N}_i \models \varphi$. Since $\mathfrak{M}_i$ extends $\mathfrak{N}_i$, it follows by monotonicity that $\mathfrak{M}_i \models \varphi$. Now let $\mathfrak{N}_k = (\mathfrak{M}_{k-1}, (E_k, E_k^{\#}))$; since $k > i$, $\mathfrak{N}_k$ is an expansion of $\mathfrak{M}_i$ and therefore φ holds in $\mathfrak{N}_k$ since it holds in $\mathfrak{M}_i$. And since $\mathfrak{N}_k$ is a fixed point for T_k, $T_k(\text{`}\varphi\text{'})$ also holds in $\mathfrak{N}_k$ and hence in $\mathfrak{M}^+$ also.

Next, let $\mathfrak{M}^+$ be any expansion of $\mathfrak{M}$ that satisfies the antecedent of Lemma 2, and let $\mathfrak{M}_n$ be its $\mathcal{L}_n$-reduct. To show that $\mathfrak{M}^+$ is a superstructure, it suffices to show that for all $n > 0$ and all φ, $(\mathfrak{M}_{n-1}, (E_n, E_n^{\#})) \models \varphi$ iff $\varphi \in E_n$ and $(\mathfrak{M}_{n-1}, (E_n, E_n^{\#})) \nvDash \varphi$ iff $\varphi \in E_n^{\#}$, where as usual E_n is T_n's extenson in $\mathfrak{M}^+$. Equivalently, letting $\mathfrak{N}$ be the model $(\mathfrak{M}_{n-1}, (E_n, E_n^{\#}))$, it suffices to show that $\mathfrak{N} \models \varphi$ iff $\mathfrak{M}^+ \models T_n(\text{`}\varphi\text{'})$ and $\mathfrak{N} \models \neg\varphi$ iff $\mathfrak{M}^+ \models T_n(\text{`}\neg\varphi\text{'})$.

These biconditionals both follow if we can show that

(a) if $\mathfrak{N} \models \varphi$ or $\mathfrak{N} \models \neg\varphi$, then $\mathfrak{M}^+ \models R_n(\text{`}\varphi\text{'})$

and

(b) if $\mathfrak{M}^+ \models R_n(\text{`}\varphi\text{'})$, then $\mathfrak{N} \models \varphi$ iff $\mathfrak{M}^+ \models \varphi$.

To prove the first biconditional from (a) and (b), suppose first that $\mathfrak{N} \models \varphi$; then by (a) and (b), $\mathfrak{M}^+ \models R_n(\text{`}\varphi\text{'})$ and $\mathfrak{M}^+ \models \varphi$; by Lemma 2(a), $\mathfrak{M}^+ \models \varphi$ iff $\mathfrak{M}^+ \models T_n(\text{`}\varphi\text{'})$. Conversely, suppose that $\mathfrak{M}^+ \models T_n(\text{`}\varphi\text{'})$; then $\mathfrak{M}^+ \models R_n(\text{`}\varphi\text{'})$, so $\mathfrak{M}^+ \models \varphi$ by Lemma 2(a) and $\mathfrak{N} \models \varphi$ by (b). This establishes the first biconditional, and the second biconditional is just a special case of the first.

We prove (a) and (b) by induction on φ's complexity. The proof is fairly straightforward, and we will only do the atomic and negation cases.

(i) φ atomic. When φ is a sentence of $\mathcal{L}_{n-1}$, the consequents of (a) and (b) hold trivially, so we may assume φ is $T_k(\text{`}\psi\text{'})$ for $k \geq n$; and when $k > n$ the antecedents of (a) and (b) fail trivially, so we may assume that in fact $k = n$. In this case (b)'s consequent is trivial; and (a)'s consequent follows by axioms (6) and (1).

(ii) $\varphi = \neg\psi$. If the antecedent of (a) holds then $\mathfrak{N} \models \psi$ or $\mathfrak{N} \models \neg\psi$, so by the induction hypothesis we have $\mathfrak{M}^+ \models R_n(\text{`}\psi\text{'})$, and hence $\mathfrak{M}^+ \models R_n(\text{`}\neg\psi\text{'})$ by axiom (2), so (a)'s consequent holds. As for (b), if $\mathfrak{M}^+ \models R_n(\text{`}\neg\psi\text{'})$ then $\mathfrak{M}^+ \models R_n(\text{`}\psi\text{'})$ since the sentences of Lemma 1 hold in $\mathfrak{M}^+$, so $\mathfrak{N} \models \psi$ iff $\mathfrak{M}^+ \models \psi$ (by the induction hypothesis) and therefore $\mathfrak{N} \models \neg\psi$ iff $\mathfrak{M}^+ \models \neg\psi$. □

THEOREM 4. *The grounded superstructure over $\mathfrak{M}$ is a Burge model.*

PROOF. Let $\mathfrak{M}^+$ be the grounded superstructure over $\mathfrak{M}$. C3$^-$ holds in $\mathfrak{M}^+$ by Theorem 3, so we only need to show that (5) holds in $\mathfrak{M}^+$. Let E_n be T_n's extension in $\mathfrak{M}^+$, and let S be any set that satisfies $\Phi_i(X)$ in $\mathfrak{M}^+$; we must show that $E_n \cup E_n^{\#} - \{\text{nonsentences}\} \subseteq S$. If X and Y are disjoint subsets of $\mathfrak{M}$'s domain such that $(\mathfrak{M}_{n-1}, (X, Y))$ is sound, let $X^\dagger = \{\varphi : (\mathfrak{M}_{n-1}, (X, Y)) \models \varphi\}$ and $Y^\dagger = \{\varphi : (\mathfrak{M}_{n-1}, (X, Y)) \nvDash \varphi\} \cup \{\text{nonsentences}\}$. Since $(E_n, E_n^{\#})$ is obtained from $(\emptyset, \emptyset)$ by repeated applications of the operation $(X, Y) \mapsto (X^\dagger, Y^\dagger)$, it suffices to show that whenever (X, Y) is a sound point of this operation,

(*) If $X \cup Y - \{\text{nonsentences}\} \subseteq S$ then $X^\dagger \cup Y^\dagger - \{\text{nonsentences}\} \subseteq S$.

(*) is proved by a routine induction on the complexity of sentences. □

The proof of Theorem 4 gives us a more general result:

COROLLARY. *A superstructure $\mathfrak{M}^+$ over $\mathfrak{M}$ is a Burge model provided that for each n, $(\mathfrak{M}_{n-1}, (E_n, E_n^{\#}))$ is the least fixed point model over some sound point $(\mathfrak{M}_{n-1}, (E, A))$ such that $E \cup A - \{$nonsentences$\}$ is contained in any S that satisfies $\Phi_i(X)$ in $\mathfrak{M}^+$.*

PROOF. Clearly the proof of Theorem 4 goes through when we replace the assumption that $(E_n, E_n^{\#})$ is the least fixed point over $(\emptyset, \emptyset)$ with the assumption that it is the least fixed point over some (E, A) with the stated properties. □

There are other superstructures besides the grounded one, and some, but not all, are Burge models. First, let's look at an ungrounded Burge model. Let $\mathfrak{M}^+$ be the unique superstructure over $\mathfrak{M}$ such that (a) for $n > 1$, $(\mathfrak{M}_{n-1}, (E_n, E_n^{\#}))$ is the least T_n-fixed point over $\mathfrak{M}_{n-1}$, and (b) $(\mathfrak{M}_0, (E_1, E_1^{\#}))$ is the least T_1- fixed point extending the sound point $(\mathfrak{M}_0, (\{\forall x\, x = x \rightarrow T_1(t)\}, \emptyset))$, where t is a term denoting $\forall x\, x = x \rightarrow T_1(t)$. It follows from the corollary to Theorem 4 that $\mathfrak{M}^+$ is a Burge model, for suppse S satisfies $\Phi_1(X)$ in $\mathfrak{M}^+$: then since $\mathfrak{M}^+ \models T_1(t)$, it follows by axiom (3) that $\forall x\, x = x \rightarrow T_1(t)$ belongs to S.

A minor variation on $\mathfrak{M}^+$'s construction yields an ungrounded superstructure that is not a Burge model. Let $\mathfrak{M}^*$ be just like $\mathfrak{M}^+$ except that now $(\mathfrak{M}_0, (E_1, E_1^{\#}))$ is the least T_1-fixed point extending $(\mathfrak{M}_0, (\{T_1(s)\}, \emptyset))$, where s is a term denoting $T_1(s)$. $T_1(s)$ belongs to E_1 but not to the smallest set S_1 satisfying $\Phi_1(X)$ in $\mathfrak{M}^*$, so $\mathfrak{M}^*$ is not a Burge model. (Intuitively, $T_1(s) \notin S_1$ because the only way $T_1(s)$ could get into S_1 is via axiom (1), but that would require the denotation of s, namely $T_1(s)$ itself, to be in S_1 already. The idea can be made rigorous using the technique of Lemma 1.) The fact that $\mathfrak{M}^*$ but not $\mathfrak{M}^+$ is a Burge model strongly suggests that the Burge models do not form an interesting class.

It is clear that the intention behind the minimality condition (5) is for the rooted$_n$ sentences to be the sentences that are grounded over $\mathfrak{M}_{n-1}$. As we have seen, however, (5) does not quite say this, since it allows the truth-teller $\forall x\, x = x \rightarrow T_1(t)$ to be rooted$_1$. A better approach is to require the set of true$_n$ sentences to be the smallest set of sentences satisfying C3$^-$. Some care must be taken in stating this requirement, however. We might, for example, try to define the Burge expansion of $\mathfrak{M}$ as the

$\leq$-least[23] expansion of $\mathfrak{M}$ to $\mathcal{L}^+$ that satisfies C3$^-$; but this wounln't work, because there is no such expansion of $\mathfrak{M}$. Indeed, any two expansions of $\mathfrak{M}$ to $\mathcal{L}^+$ that satisfy C3$^-$ are $\leq$-incomparable. To see this, let $\mathfrak{M}'$, $\mathfrak{M}''$ be two such expansions, and suppose for contradiction that $\mathfrak{M}' \leq \mathfrak{M}''$. Since $\mathfrak{M}'$ and $\mathfrak{M}''$ are distinct, we must have $\mathfrak{M}'' \models T_n(\text{`}\varphi\text{'})$ and $\mathfrak{M}' \models \neg T_n(\text{`}\varphi\text{'})$ for some φ and n. But then by (0) and (7) we have $\mathfrak{M}' \models T_{n+1}(\text{`}\neg T_n(\text{`}\varphi\text{'})\text{'})$ but not $\mathfrak{M}'' \models T_{n+1}(\text{`}\neg T_n(\text{`}\varphi\text{'})\text{'})$, contradicting $\mathfrak{M}' \leq \mathfrak{M}''$.

The correct approach is not to require the expansion of $\mathfrak{M}$ to be minimal when taken as a whole, but to require instead that the extension of each T_n be minimal when we hold fixed the extensions of all the T_m with $m < n$. Thus we have the following:

THEOREM 5. *The grounded superstructure over $\mathfrak{M}$ is the unique expansion $\mathfrak{M}^+$ of $\mathfrak{M}$ to $\mathcal{L}^+$ in which* C3$^-$ *holds and that satisfies*

(#) *For all $n > 0$, the extension of T_n in $\mathfrak{M}^+$ is the smallest set S such that $(\mathfrak{M}_{n-1}, S)$ satisfies those instances of the axioms of* C3$^-$ *in which all subscripts occurring outside quotes are $\leq n$.*

PROOF. Let $\mathfrak{M}^+$ be the grounded superstructure over $\mathfrak{M}$; for each n, let E_n be T_n's extension in $\mathfrak{M}^+$. Since C3$^-$ holds in $\mathfrak{M}^+$ by Theorem 4, we need only show that $\mathfrak{M}^+$ satisfies (#), since clearly any expansion of $\mathfrak{M}$ satisfying (#) is unique. Fix n, and let S be any set such that $(\mathfrak{M}_{n-1}, S)$ satisfies the instances of the axioms of C3$^-$ that are sentences of $\mathcal{L}_n$. We will show that $E_n \subseteq S$. Similarly to Theorem 4, it suffices to show that $X^\dagger \subseteq S$ whenever $X \subseteq S$ and $(\mathfrak{M}_{n-1}, (X, X^\#))$ is sound, where $X^\dagger = \{\varphi : (\mathfrak{M}_{n-1}, (X, X^\#)) \models \varphi\}$. Let X be as required, and let $\mathfrak{N}$ be the model $(\mathfrak{M}_{n-1}, (X, X^\#))$; we will show by induction on φ's complexity that if $\mathfrak{N} \models \varphi$ then $\varphi \in S$ and if $\mathfrak{N} \models \varphi$ then $\neg\varphi \in S$.

φ atomic: The only nontrivial case is when φ is the sentence $T_n(\text{`}\psi\text{'})$. Suppose first that $\mathfrak{N} \models \varphi$. Then $\psi \in X$ since $\mathfrak{N}$ is sound, and therefore $\psi \in S$; the following

[23]Where $\mathfrak{M} \leq \mathfrak{N}$ just in case each T_i's extension in $\mathfrak{M}$ is contained in its extension in $\mathfrak{N}$, and $\mathfrak{M}$ is otherwise identical to $\mathfrak{N}$.

sentences are therefore true in $(\mathfrak{M}_{n-1}, S)$: $T_n(\text{`}\psi\text{'})$, $R_n(\text{`}\psi\text{'})$, $R_n(\text{`}T_n(\text{`}\psi\text{'})\text{'})$ (by (1)), and $T_n(\text{`}T_n(\text{`}\psi\text{'})\text{'}) \leftrightarrow T_n(\text{`}\psi\text{'})$ (by (10)). It follows that $(\mathfrak{M}_{n-1}, S) \models T_n(\text{`}T_n(\text{`}\psi\text{'})\text{'})$, i.e., $\varphi \in S$. Next suppose that $\mathfrak{N} \neq \varphi$. Then $\psi \in X^{\#}$, so $\neg\psi \in X \subseteq S$. The following are therefore true in $(\mathfrak{M}_{n-1}, S)$: $T_n(\text{`}\neg\psi\text{'})$, $R_n(\text{`}\psi\text{'})$, $R_n(\text{`}T_n(\text{`}\psi\text{'})\text{'})$, $R_n(\text{`}\neg T_n(\text{`}\psi\text{'})\text{'})$, $T_n(\text{`}\neg T_n(\text{`}\psi\text{'})\text{'}) \leftrightarrow \neg T_n(\text{`}T_n(\text{`}\psi\text{'})\text{'})$ (by (7)), $T_n(\text{`}T_n(\text{`}\psi\text{'})\text{'}) \leftrightarrow T_n(\text{`}\psi\text{'})$ (by (10)), and $\neg T_n(\text{`}\psi\text{'})$ (by (7), since $T_n(\text{`}\neg\psi\text{'})$ holds in $(\mathfrak{M}_{n-1}, S)$). It follows that $T_n(\text{`}\neg T_n(\text{`}\psi\text{'})\text{'})$ holds in $(\mathfrak{M}_{n-1}, S)$, i.e., $\neg\varphi \in S$.

$\varphi = \neg\psi$: Trivial.

$\varphi = \psi \rightarrow \chi$: Suppose $\mathfrak{N} \models \varphi$. Then $\mathfrak{N} \neq \psi$ or $\mathfrak{N} \models \chi$; by the induction hypothesis, either $\neg\psi \in S$ or $\chi \in S$. By axiom (3), then, $(\mathfrak{M}_{n-1}, S) \models R_n(\text{`}\varphi\text{'})$. Since $(\mathfrak{M}_{n-1}, S) \models T_n(\text{`}\neg\psi\text{'}) \rightarrow \neg T_n(\text{`}\psi\text{'})$ by (7), we have $(\mathfrak{M}_{n-1}, S) \models \neg T_n(\text{`}\psi\text{'}) \vee T_n(\text{`}\chi\text{'})$, i.e., $(\mathfrak{M}_{n-1}, S) \models T_n(\text{`}\psi\text{'}) \rightarrow T_n(\text{`}\chi\text{'})$; by (8), then, we have $(\mathfrak{M}_{n-1}, S) \models T_n(\text{`}\psi \rightarrow \chi\text{'})$. Similar reasoning shows that if $\mathfrak{N} \neq \varphi$ then $\neg\varphi \in S$.

$\varphi = \forall x\, \psi$: This is similar to the previous case. □

A.2. Construction C4. Burge remarks (in [**Bur79**, p. 205]) that C3 can be further liberalized by allowing the subscripts of (1) that occur inside quotes to be $\geq i$. In terms of our formulation, that would mean changing (1) to

$$(1^*) \qquad R_i(t) \rightarrow R_i(\text{`}T_j(t)\text{'}).$$

Call the resulting system C4. C4 has such consequences as

$$2 + 2 = 4 \rightarrow T_i(\text{`}T_{i+3}(\text{`}2 + 2 = 4\text{'})\text{'}).$$

Whether it also has consequences of the form $\varphi \wedge \neg\varphi$ is a matter that Burge unfortunately does not discuss. What follows is a proof of C4's consistency.

Our construction will proceed more or less as follows. Once we have determined the extensions of $T_1 \ldots T_{n-1}$, we first assign extension-antiextension pairs to *all* the other predicates T_n, T_{n+1},..., so that each one becomes a truth predicate for the

entire language $\mathcal{L}^+$, using the usual inductive procedure. Then we "close off" the extension of T_n (but not of $T_{n+1}, \ldots$), just as before. The resulting model will satisfy C4, and will be the unique model satisfying a suitable modification of C4.

Now for the details. As before, we assume a fixed classical model $\mathfrak{M}$ for $\mathcal{L}$. If $\mathfrak{M}'$ is a partial model for $\mathcal{L}^+$ which is an expansion of $\mathfrak{M}$, let $\Gamma_n\mathfrak{M}'$ be the model that is just like $\mathfrak{M}'$, except that for all $k \geq n$ it assigns T_k the extension $\{\varphi : \mathfrak{M}' \models \varphi\}$ and antiextension $\{\varphi : \mathfrak{M}' \nvDash \varphi\} \cup \{\text{nonsentences}\}$. Obviously Γ_n is a monotone operator. Define Γ_n^α for ordinals α by $\Gamma_n^0\mathfrak{M}' = \mathfrak{M}'$, $\Gamma_n^{\alpha+1}\mathfrak{M}' = \Gamma_n\Gamma_n^\alpha\mathfrak{M}'$, $\Gamma_n^\lambda\mathfrak{M}' = \bigcup_{\alpha<\lambda}\Gamma_n^\alpha\mathfrak{M}'$ for λ a limit or ∞. Notice that if $\mathfrak{M}'$ is not Γ_n-sound, there is no guarantee that $\Gamma_n^\alpha\mathfrak{M}'$ is defined for all α; on the other hand, if $\mathfrak{M}'$ is Γ_n-sound then $\Gamma_n^\infty\mathfrak{M}'$ is the least Γ_n-fixed point extending $\mathfrak{M}'$. Define $\mathrm{Cl}_n\mathfrak{M}'$ to be the result of "closing off" the predicate T_n in $\mathfrak{M}'$, i.e., $\mathrm{Cl}_n\mathfrak{M}'$ is just like $\mathfrak{M}'$ except that where T_n's interpretation in $\mathfrak{M}'$ is (E, A), its interpretation in $\mathrm{Cl}_n\mathfrak{M}'$ is $(E, -E)$. We define $\mathfrak{M}_n$ inductively as follows: $\mathfrak{M}_0 = \mathfrak{M}$, $\mathfrak{M}_{n+1} = \mathrm{Cl}_{n+1}\Gamma_{n+1}^\infty\mathfrak{M}_n$. $\mathfrak{M}_\omega$ is the limit this sequence approaches, i.e., it is the expansion of $\mathfrak{M}$ to $\mathcal{L}^+$ that assigns T_n the same interpretation that $\mathfrak{M}_n$ does. (Thus, $\mathfrak{M}_n$ is total up through T_n and partial thereafter, and $\mathfrak{M}_\omega$ is a total model.) We will prove that $\mathfrak{M}_\omega$ satisfies C4.

In order for $\mathfrak{M}_\omega$ to be well-defined, $\mathfrak{M}_{n-1}$must always be a Γ_n-sound point; so we will begin by showing that this is the case.

LEMMA 3. *$\mathfrak{M}_{n-1}$ is Γ_n-sound.*

PROOF. We assume that $\mathfrak{M}_i$ is well-defined for $i < n$ and $\mathfrak{M}_{i-1}$ is Γ_i-sound for $0 < i < n$, and prove that $\mathfrak{M}_{n-1}$ is Γ_n-sound. That is, we will show that $\mathfrak{M}_{n-1} \subseteq \Gamma_n\mathfrak{M}_{n-1}$, which is just to say that for each sentence φ of $\mathcal{L}^+$ and each k, if $\mathfrak{M}_{n-1} \models T_k(\text{`}\varphi\text{'})$ then $\Gamma_n\mathfrak{M}_{n-1} \models T_k(\text{`}\varphi\text{'})$ and if $\mathfrak{M}_{n-1} \models \neg T_k(\text{`}\varphi\text{'})$ then $\Gamma_n\mathfrak{M}_{n-1} \models \neg T_k(\text{`}\varphi\text{'})$. This holds trivially for $k < n$ by the definition of Γ_n, so we may assume that $k \geq n$. Now for such k, $\Gamma_n\mathfrak{M}_{n-1} \models T_k(\text{`}\varphi\text{'})$ iff $\mathfrak{M}_{n-1} \models \varphi$ and $\Gamma_n\mathfrak{M}_{n-1} \models \neg T_k(\text{`}\varphi\text{'})$ iff $\mathfrak{M}_{n-1} \models \neg\varphi$; so what we have to show is that $\mathfrak{M}_{n-1} \models T_k(\text{`}\varphi\text{'})$ implies $\mathfrak{M}_{n-1} \models \varphi$

and $\mathfrak{M}_{n-1} \models \neg T_k(\ulcorner\varphi\urcorner)$ implies $\mathfrak{M}_{n-1} \models \neg\varphi$. If $n = 1$ then this holds vacuously, since we never have $\mathfrak{M}_0 \models T_k(\ulcorner\varphi\urcorner)$ or $\mathfrak{M}_0 \models \neg T_k(\ulcorner\varphi\urcorner)$. If $n > 1$ then $\mathfrak{M}_{n-1} = \mathrm{Cl}_{n-1}\Gamma^\infty_{n-1}\mathfrak{M}_{n-2}$, and $\mathrm{Cl}_{n-1}\Gamma^\infty_{n-1}\mathfrak{M}_{n-2} \models \neg T_k(\ulcorner\varphi\urcorner)$ iff $\Gamma^\infty_{n-1}\mathfrak{M}_{n-2} \models \neg T_k(\ulcorner\varphi\urcorner)$ (since $k \neq n-1$) iff $\Gamma^\infty_{n-1}\mathfrak{M}_{n-2} \models \neg\varphi$; and clearly $\mathfrak{M}_{n-1} \models T_k(\ulcorner\varphi\urcorner)$ iff $\Gamma^\infty_{n-1}\mathfrak{M}_{n-2} \models \varphi$. But since $\Gamma^\infty_{n-1}\mathfrak{M}_{n-2} \subseteq \mathrm{Cl}_{n-1}\Gamma^\infty_{n-1}\mathfrak{M}_{n-2} = \mathfrak{M}_{n-1}$ and by the monotonicity of the strong Kleene scheme, $\Gamma^\infty_{n-1}\mathfrak{M}_{n-2} \models \varphi$ implies $\mathfrak{M}_{n-1} \models \varphi$ and $\Gamma^\infty_{n-1}\mathfrak{M}_{n-2} \models \neg\varphi$ implies $\mathfrak{M}_{n-1} \models \neg\varphi$. □

Next we must verify that the axioms of C4 hold in $\mathfrak{M}_\omega$. We will only prove this for (1*) and (5); the other cases are completely straightforward.

THEOREM 6. *The axioms of* C4 *hold in* $\mathfrak{M}_\omega$.

PROOF. (1*) holds in $\mathfrak{M}_\omega$: assume $\mathfrak{M}_\omega \models R_i(t)$, where t denotes φ, to show that $\mathfrak{M}_\omega \models R_i(\ulcorner T_j(t)\urcorner)$. We may assume that $j \geq i$, since it's trivial when $j < i$. $\mathfrak{M}_i \models R_i(t)$, i.e., $\mathfrak{M}_i \models T_i(\ulcorner\varphi\urcorner)$ or $\mathfrak{M}_i \models T_i(\ulcorner\neg\varphi\urcorner)$. If $\mathfrak{M}_i \models T_i(\ulcorner\varphi\urcorner)$, then $\Gamma^\infty_i\mathfrak{M}_{i-1} \models T_i(\ulcorner\varphi\urcorner)$; but since $\Gamma^\infty_i\mathfrak{M}_{i-1}$ is a Γ_i-fixed point, $\Gamma^\infty_i\mathfrak{M}_{i-1} \models T_i(\ulcorner\varphi\urcorner)$ iff $\Gamma^\infty_i\mathfrak{M}_{i-1} \models \varphi$ iff $\Gamma^\infty_i\mathfrak{M}_{i-1} \models T_j(t)$, iff $\Gamma^\infty_i\mathfrak{M}_{i-1} \models T_i(\ulcorner T_j(t)\urcorner)$; hence, $\Gamma^\infty_i\mathfrak{M}_{i-1} \models T_i(\ulcorner T_j(t)\urcorner)$ and $\mathfrak{M}_i \models T_i(\ulcorner T_j(t)\urcorner)$. Similar reasoning applies if $\mathfrak{M}_i \models T_i(\ulcorner\neg\varphi\urcorner)$.

(5) holds in $\mathfrak{M}_\omega$: We assume that (5) is formalized as a second order scheme $\forall x\left[R_i(x) \to \forall x\left(\Psi_i(X) \to X(x)\right)\right]$ analogous to the (5) of C3. For each i, let S_i be the least set that satisfies $\Psi_i(X)$ in $\mathfrak{M}_\omega$. To prove that (5) holds in $\mathfrak{M}_\omega$ it suffices to show that $\{\varphi : \mathfrak{M}_\omega \models R_i(\ulcorner\varphi\urcorner)\} \subseteq S_i$ for all i. We will show this by proving by induction that for all α and n,

(**) for all $i > 0$ and all φ, if $\Gamma^\alpha_{n+1}\mathfrak{M}_n \models R_i(\ulcorner\varphi\urcorner)$ then $\varphi \in S_i$.

(i) (**) holds for $\alpha = n = 0$ trivially since we never have $\mathfrak{M}_0 \models R_i(\ulcorner\varphi\urcorner)$ for any φ.

(ii) Fix n and assume that for all β, (**) holds for β, n to show that it holds for $0, n+1$. Then we have $\{\varphi : \Gamma^\infty_{n+1}\mathfrak{M}_n \models R_i(\ulcorner\varphi\urcorner)\} \subseteq S_i$ for all i; but $\{\varphi : \Gamma^\infty_{n+1}\mathfrak{M}_n \models R_i(\ulcorner\varphi\urcorner)\} = \{\varphi : \mathrm{Cl}_{n+1}\Gamma^\infty_{n+1}\mathfrak{M}_n \models R_i(\ulcorner\varphi\urcorner)\} = \{\varphi : \mathfrak{M}_{n+1} \models R_i(\ulcorner\varphi\urcorner)\}$.

(iii) Fix n and $\alpha > 0$ and assume that for all $\beta < \alpha$, (**) holds for β, n to show that it holds for α, n. What we want to show is that $\{\varphi : \Gamma^{\alpha}_{n+1}\mathfrak{M}_n R_i(\text{`}\varphi\text{'})\} \subseteq S_i$. This obviously holds if $\alpha$ is a limit, so we may assume that $\alpha$ is a successor, say $\alpha = \beta + 1$. If $i \leq n$ then $\{\varphi : \Gamma^{\alpha}_{n+1}\mathfrak{M}_n \models R_i(\text{`}\varphi\text{'})\} = \{\varphi : \Gamma^{\beta}_{n+1}\mathfrak{M}_n \models R_i(\text{`}\varphi\text{'})\}$ for all $\varphi$, so assume $i > n$. Now since $\Gamma^{\alpha}_{n+1}\mathfrak{M}_n \models T_i(\text{`}\varphi\text{'})$ iff $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \varphi$, it suffices to show for all φ that if $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \varphi$ then $\varphi \in S_i$ and if $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \neg\varphi$ then $\varphi \in S_i$. We show this by induction on φ's complexity.

φ atomic: If φ is a sentence of $\mathcal{L}_n$, then $\varphi \in S_i$ by axiom (0). So suppose $\varphi = T_j(t)$ for some $j > n$, where t denotes ψ. First, suppose $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models T_j(t)$. $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models T_j(t)$ iff $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models T_j(\text{`}\psi\text{'})$ iff $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models T_i(\text{`}\psi\text{'})$ (since $i, j > n$), so by the induction hypothesis, $\psi \in S_i$. Since S_i satisfies (1*), it follows that $T_j(t) \in S_i$.

Next, assume $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \neg T_j(\text{`}\psi\text{'})$. $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \neg T_j(\text{`}\psi\text{'})$ iff $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models T_j(\text{`}\neg\psi\text{'})$ (because $j > n$), so by the foregoing $\neg\psi \in S_i$. But clearly we can only have $\neg\psi \in S_i$ if $\psi \in S_i$, so again $\psi \in S_i$.

$\varphi = \neg\psi$: Trivial.

$\varphi = (\psi \to \chi)$: If $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \neg\varphi$, then $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \psi$ and $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \neg\chi$, so $\psi, \chi \in S_i$ by the induction hypothesis; so $\psi \to \chi \in S_i$ by axiom (3). If $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \varphi$, then either $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \neg\psi$ or $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \models \chi$, and so either $\mathfrak{M}_\omega \models \neg\psi$ or $\mathfrak{M}_\omega \models \chi$ (by monotonicity and since $\Gamma^{\beta}_{n+1}\mathfrak{M}_n \subseteq \mathfrak{M}_\omega$); then by (3) again $\psi \to \chi \in S_i$.

$\varphi = \forall x\, \psi$: This is similar to the previous case. □

As with C3, C4's axiom (5) probably doesn't quite capture what Burge had in mind, and should probably be rephrased. If we amend C4 the same way we amended C3, then $\mathfrak{M}_\omega$ turns out to be the resulting construction's unique model. In what follows, C4$^-$ consists of the axioms of C4 other than (5).

THEOREM 7. *$\mathfrak{M}_\omega$ is the unique classical expansion $\mathfrak{M}'$ of $\mathfrak{M}$ to $\mathcal{L}^+$ such that for each $n > 0$,*

(##) *The extension of* T_n *in* $\mathfrak{M}'$ *is the smallest set* S *such that there exists an expansion of* $\mathfrak{M}_{n-1}$ *to* $\mathcal{L}^+$ *in which* T_n*'s extension is* S *and that satisfies the instances of the axioms of* C4$^-$.

($\mathfrak{M}_{n-1}$ *is* $\mathfrak{M}'$*'s* $\mathcal{L}_{n-1}$*-reduct*).

Proof. As with C3, if such a model exists then it is unique, so we need only verify that $\mathfrak{M}_\omega$ satisfies (##) for each $n > 0$. Fix n; obviously there is an expansion $\mathfrak{M}^*$ of $\mathfrak{M}_{n-1}$ that satisfies C4$^-$, since we may take $\mathfrak{M}^* = \mathfrak{M}_\omega$. So let S and $\mathfrak{M}^*$ be a set and an expanson of $\mathfrak{M}_{n-1}$ to $\mathcal{L}^+$, respectively, such that $\mathfrak{M}^*$ satisfies C4$^-$ and S is T_n's extension in $\mathfrak{M}^*$: we need to show that $E_n \subseteq S$, where $E_n = T_n$'s extension in $\mathfrak{M}_\omega$. For any disjoint X and Y let $(\mathfrak{M}_{n-1}, (X,Y), (X,Y), \ldots)$ be the expansion of $\mathfrak{M}_{n-1}$ to $\mathcal{L}^+$ in which each T_k for $k \geq n$ is interpreted as (X,Y); similarly to Theorem 5, it suffices to show that if $X \subseteq S$ and $Y = X^\#$ then $X^* \subseteq S$, where $(\mathfrak{M}_{n-1}, (X^*,Y^*), (X^*,Y^*), \ldots) = \Gamma_n(\mathfrak{M}_{n-1}, (X,Y), (X,Y), \ldots)$. That is, for all φ we must show that if $(\mathfrak{M}_{n-1}, (X,Y), (X,Y), \ldots) \models \varphi$ then $\varphi \in S$. The proof of this is very similar to the proof of the corresponding fact in Theorem 5, and we will only do the case of φ atomic here.

φ atomic. The only interesting subcase is $\varphi = T_k(\text{`}\psi\text{'})$ for $k \geq n$. Let $\mathfrak{N} = (\mathfrak{M}_{n-1}, (X,Y), (X,Y), \ldots)$; we will show that if $\mathfrak{N} \models \varphi$ then $\varphi \in S$ and if $\mathfrak{N} \models\!\!\!| \varphi$ then $\neg\varphi \in S$. First suppose $\mathfrak{N} \models \varphi$. Then $\psi \in X \subseteq S$, so $\mathfrak{M}^* \models T_n(\text{`}\psi\text{'})$. The following then hold in $\mathfrak{M}^*$: $T_k(\text{`}\psi\text{'})$ (by (11)), $R_n(\text{`}\psi\text{'})$, $R_n(\text{`}T_k(\text{`}\psi\text{'})\text{'})$ (by (1*)), $T_n(\text{`}T_k(\text{`}\psi\text{'})\text{'}) \leftrightarrow T_k(\text{`}\psi\text{'})$ (by (10)), $T_n(\text{`}T_k(\text{`}\psi\text{'})\text{'})$: i.e., $T_k(\text{`}\psi\text{'}) \in S$. Next suppose $\mathfrak{N} \models\!\!\!| \varphi$. Then $\psi \in Y$ and $\neg\psi \in X \subseteq S$, so the following are true in $\mathfrak{M}^*$: $T_n(\text{`}\neg\psi\text{'})$, $R_n(\text{`}\psi\text{'})$, $R_n(\text{`}\neg\psi\text{'})$, $R_n(\text{`}T_k(\text{`}\psi\text{'})\text{'})$, $R_n(\text{`}\neg T_k(\text{`}\psi\text{'})\text{'})$, $T_n(\text{`}\neg T_k(\text{`}\psi\text{'})\text{'}) \leftrightarrow \neg T_n(\text{`}T_k(\text{`}\psi\text{'})\text{'})$ (by (7)), $T_n(\text{`}T_k(\text{`}\psi\text{'})\text{'}) \leftrightarrow T_k(\text{`}\psi\text{'})$ (by (10)), $T_k(\text{`}\neg\psi\text{'})$ (by (11)), $\neg T_k(\text{`}\psi\text{'})$ (by (7)), and therefore $T_n(\text{`}\neg T_k(\text{`}\psi\text{'})\text{'})$. □

CHAPTER 5

The Inconsistency Theory

1. Where We Are

The last two chapters were intended to serve two purposes. First, in raising problems for several existing views, they make any alternative, such as my theory, at least worth considering and perhaps even more plausible than it would otherwise be. Second, we will see that they support my view more directly by raising doubts about the possibility of any understanding of truth that is both consistent and intuitive.

What drives the negative results of the last two chapters is the tendency of disquotational principles (like (T) or RT_1^*–RT_4^*) to generate contradictions. These contradictions are generated in two different ways. First, the rules often yield contradictions by themselves in the presence of suitable logics.[1] Second, they can be used to undermine positive proposals about the Liar. Let's briefly recapitulate our main results.

• The rules RT_1^*–RT_4^* are inconsistent in a wide variety of logics. (The same applies to (T).) While they are consistent in the logic of the strong Kleene scheme, natural languages appear to have features that require going beyond that scheme in a way that makes the rules inconsistent again.

• Consistency can be gained by severely restricting the rules to RT_1–RT_4; but this restriction is very unnatural and appears to have no basis in how we actually use the truth predicate.

• Even with this restriction in place, moreover, inconsistency is regained as soon as one makes a claim about the Liar's truth value.

[1]Strictly speaking, deriving a contradiction in these cases also requires use of an extra-logical assumption like 'S = 'S is not true''. This qualification should always be understood when I say that a rule or schema is inconsistent.

• This last point seems to be very persistent, in that analogous problems seem to arise no matter what sort of defectiveness one ascribes to the Liar.

If formal inconsistency is unavoidable, one might still argue that this is not *real* inconsistency, citing some sort of contextually determined hidden parameter. We saw that this idea gives rise to a great variety of views that I have collectively referred to as the "hierarchy approach." We also saw that, when nothing else is wrong with these views, they have the irksome feature of being unstatable by their own lights. While some philosophers go in for this sort of thing, others find it deeply problematic, and I have urged the latter point of view in this essay. And while it is probably impossible to have the last word on this issue, I intend to rest my case and move on, assuming from now on that the hierarchy approach is not viable.

This leaves us with the conclusion that, on pain of inconsistency, we must accept that truth simply does not have the disquotational properties it seems to have (although of course it may approximate them). But it is very hard to see how we can possibly accept this. Rules like (T) seem like rather obvious definitional facts, and this impression persists even when one brings problematic cases clearly into focus. To say 'the Liar is not true, but it's not true that the Liar is not true' seems rather like saying that there is an uncle who was never anyone's brother. It is certainly possible for intuitions to be incorrect, even linguistic intuitions; but our intuitions about truth seem amazingly resistant to education.

My proposal, simply put, is that the truth predicate *purports* to have disquotational properties. Put another way, the schema (T) (or the rules RT_1^*–RT_4^* or something similar) is built into the truth concept in the sense that the fact of our meaning *true* by 'true' consists simply in our having accepted (T). This view explains why (T) seems so inevitable, but it does so without itself suffering inconsistency. My proposal provides a clear sense in which the *concept* of truth, not just this or that

theory of truth, is inconsistent; because of this, I call it the "inconsistency theory of truth."[2]

2. What Are Disquotational Properties?

Before defending the claim that we have accepted a disquotational rule, though, we should first settle the matter of exactly what rule this is. A straightforward answer is simply the schema (T). This is the answer I will eventually favor, but the matter needs closer examination.

The reason has to do with truth value gaps. Gaps unquestionably occur in natural language, even though it is questionable whether they play an important role in explaining the Liar. And if we take the most popular approach to handling gaps, the strong Kleene scheme, we will find that (T) is unacceptable for reasons having nothing to do with the paradoxes. Specifically, in the strong Kleene scheme a biconditional is always gappy if either of its sides is gappy; thus if S is a gappy sentence, the corresponding instance of (T), "S' is true iff S', is gappy. So accepting (T) commits us to all the gappy instances of (T).

This assumes that 'iff' is interpreted in the standard way; and fortunately there is an alternative interpretation available. As was mentioned in Chapter 3, some natural language sentences express penumbral connections between vague predicates. There we saw how some such connections might be expressed by means of a conditional; others are most naturally expressed via a biconditional, e.g.,

(1) Alma is having a stroll iff she is having a leisurely walk.

On the standard interpretation of the biconditional (and assuming the strong Kleene scheme), (1) would be gappy whenever Alma's activity is on the borderline between stroll and non-stroll. Yet (1) seems true. This suggests that 'iff' should be interpreted

[2]My view is actually not a theory of truth in the same way that, say, Burge's or McGee's is, since I am not giving an account of which sentences or propositions are true and which are not true, but rather describing our concept of truth. So 'inconsistency theory of the concept of truth' would be more accurate, if rather clumsy.

in some nonstandard way in this and similar cases; let us use '$\equiv$' to represent this interpretation of 'iff', and reserve '$\leftrightarrow$' for the standard interpretation.[3]

My proposal, then, is that the disquotational rule we accept as part of the meaning of 'true' is

(T) $$\text{`}A\text{' is true } \equiv A$$

This avoids the above problem, since $\varphi \equiv \psi$ may well be true even when φ and ψ are gappy. This formulation of (T) is particularly natural if one thinks that 'true' is a vague predicate as suggested in Chapter 3—or at least that 'true' purports to be a vague predicate—since then the truth schema is naturally seen as expressing a penumbral connection.

As to whether (T) is consistent, let us first note that if natural language contains $\equiv$, it presumably also contains the connective $\supset$ of Chapter 3. (Indeed, $\varphi \equiv \psi$ might well be seen as an abbreviation for $(\varphi \supset \psi) \wedge (\psi \supset \varphi)$.) As we noted there (and in Chapter 2), the rules RT_1^* and RT_2^* yield a contradiction in the presence of this connective (via Curry's paradox), provided it satisfies the rules of conditional proof and *modus ponens*. And RT_1^* and RT_2^* are derivable from (T), provided $\equiv$ satisfies the rule

($\mathrm{MP}_{\equiv}$) $$\frac{\varphi, \varphi \equiv \psi}{\psi} \qquad \frac{\psi, \varphi \equiv \psi}{\varphi}$$

which any biconditional worthy of the name should certainly satisfy. And so (T) is inconsistent.

Interestingly, there is no obvious inconsistency in this instance of (T): 'The Liar is true $\equiv$ the Liar is not true'. To generate an inconsistency from this instance, we need to make some assumption about the Liar's truth value. I have already argued that a principled reticence about the Liar's truth value cannot ultimately avoid the

[3] $\equiv$ might be interpreted truth functionally, say, by a truth table just like the one for $\leftrightarrow$ except that $\varphi \equiv \psi$ is true whenever φ and ψ are gappy. Or it might be given some non-truth-functional interpretation. The precise details of how $\equiv$ is interpreted are not important for our purposes.

strengthened Liar problem; nonetheless, as far as the consistency of (T) is concerned, Curry's paradox seems to be of more direct relevance than the traditional form of the Liar.[4]

An alternative to (T) is the set of rules RT^*_1–RT^*_4. But this is not really an alternative, since RT^*_1–RT^*_4 are equivalent to (T) if we assume that $\varphi \equiv \psi$ is equivalent to $(\varphi \supset \psi) \wedge (\psi \supset \varphi)$ and that $\supset$ satisfies the *modus ponens* and conditional proof rules. (The proof is simple and is left to the reader.)

Thus far I have been treating 'true' as applying to sentences; however, I believe that in ordinary English, the term applies primarily to propositions. (T) ought to be revised to reflect this; the most obvious way is as follows:

(T) The proposition that p is true $\equiv p$

Thus, finally, we arrive at our official version of (T). Now we need to ask: is this version of (T) inconsistent? The answer is yes, provided we are willing to accept some obvious claims about propositions. In particular, the argument of Curry's paradox still works with the propositional version of (T). Begin by considering the sentence

(2) If (2) expresses a true proposition, then A

(where as before A can be any sentence). We reason as follows:

(1) (2) expresses a true proposition (assumption)

(2) The proposition that if (2) expresses a true proposition then A, is true (from 1)

[4]All this assumes a strong Kleene-style handling of truth value gaps. Another approach is supervaluatonal: a sentence is regarded as true just in case it comes out true on any acceptable way of assigning truth values to gappy atomic sentences. On such an approach, matters are much simpler. (1) comes out true even when 'iff' is given its usual truth functional interpretation, since any acceptable precisification of 'stroll' and 'leisurely' and 'walk' will make it true. At the same time, any sentence of the form $\varphi \leftrightarrow \neg\varphi$ is supervaluationally false, since no way of assigning a truth value to φ can make it true. So if we take a supervaluational approach, we can simply take (T) to be "'A' is true $\leftrightarrow A$'.

(3) If (2) expresses a true proposition, then A (from 2 and (T))

(4) A (*modus ponens*, 1 and 3)

(5) If (2) expresses a true proposition, then A (conditional proof, 1–4)

(6) The proposition that if (2) expresses a true proposition then A, is true (5 and (T))

(7) 'If (2) expresses a true proposition then A' expresses the proposition that if (2) expresses a true proposition, then A

(8) 'If (2) expresses a true proposition, then A' expresses a true proposition (6 and 7)

(9) (2) expresses a true proposition (8)

(10) A (5 and 9)

The inference from 1 to 2 goes through if we assume

$$\forall x\,(\text{`}S\text{'} \text{ expresses } x \supset x \text{ is the propositin that } S)$$

and step 7 is justified by the rule

$$\frac{S}{\text{`}S\text{'} \text{ expresses a proposition}}$$

where this rule applies only to assertions. I think it's pretty clear that we must accept both principles if we accept the notion of a proposition at all.

3. What the Theory Is

In [**Tar35**], Tarski famously claimed that the semantic paradoxes show natural languages to be inconsistent.[5] Since I am claiming that, as a matter of linguistic convention, the truth predicate purports to satisfy the disquotational schema (T), and since that schema is in fact inconsistent, I am arguing for a version of Tarski's view.

[5]He takes a more qualified position in [**Tar44**]: there he claims that any *formalized* language meeting certain requirements is inconsistent, and that it is perhaps a good *guess* that the formalized languages that best represent natural ones satisfy those requirements.

While this puts me in good company, it also gives me some explaining to do. Tarski did little to explain his view, and essentially nothing to defend it. As a result, there is now a longstanding tradition of taking Tarski to task for these omissions. Many of Tarski's critics focus on the difficulty of making any sense of the view at all; the following remark by Tappenden is fairly typical:

> To put it mildly, it is not obvious what [Tarski's claim] could mean. One is left to ask rhetorically "How could a natural language be inconsistent?" The question is not asked in the manner of one who understands what it would be for a language to be inconsistent but denies that a natural language could be inconsistent in the manner recognized. Rather the question is meant to evince bafflement at how one is to understand the idea of an inconsistent language at all. [**Tap92**, p. 151]

Others are more openly hostile to the view, seeing it less as an obscure doctrine and more as a basic confusion. In this spirit, Burge writes:

> According to [Tarski's view], part of the nature of a "language" is a set of postulates that purport to be true by virtue of their meaning or are at least partially constitutive of that "language". Tarski thought that he had identified just such postulates in natural language as spawning inconsistency. But postulates are contained in theories that are promoted by people. Natural languages per se do not postulate or assert anything. What engenders paradox is a certain naive theory or conception of the natural concept of truth. It is the business of those interested in natural language to improve on it. [**Bur79**, pp. 83–4]

The reception of Tarski's view has not been wholly negative. The most notable defense of a Tarskian conception of truth is Chihara's [**Chi79**]; there he argues that the concept of truth is inherently paradoxical, roughly in the sense that the inconsistent schema (T) is part of the meaning of 'true'. This entire essay can be viewed as

an attempt to work out some of Chihara's ideas. However, most other authors who express sympathy with Tarski's view do little to clarify it, and some are even more cryptic than Tarski himself.

Now Tarski did give more of a clue as to what he meant than Tappenden's comment suggests, though many have found his clue inadequate. In [**Tar35**], Tarski explicitly states that part of what it is to specify a *formalized* language is to specify a set of *axioms*, to be thought of somewhat along the lines of analytic statements. It's not hard to guess that an inconsistent formalized language is one whose set of axioms is inconsistent, and that Tarski thinks a formalized version of a natural language would include the disquotational schema (or its instances) among its axioms, making it inconsistent in a derivative sense (see note 5). It is these axioms, or rather their informal counterparts, that Burge is referring to when he speaks of "postulates that purport to be true by virtue of their meaning."

The idea of a language containing axioms is completely innocuous if that language is a formal one; but Burge and others are right to ask what it could be for something to be an "axiom" of a *natural* language. The main problem people have with Tarski's view is that he doesn't explain, nor is it clear, what it is about the axioms that distinguishes them from any other sentences.

A natural first attempt at an answer is that the axioms are analytic; but this doesn't get us very far. The familiar, standard notion of an analytic sentence is that of a sentence that is *true* by virtue of its meaning. If this is what is meant by 'analytic', and if a natural language's axioms are its analytic sentences, then an inconsistent language is one whose analytic sentences are mutually inconsistent. But now the very claim that a language is inconsistent is itself inconsistent: if L is inconsistent, then L's axioms are all true (being analytic), and at the same time at least one of them is not true (since they are mutually inconsistent). This is the substance of Herzberger's influential critique of the notion of an inconsistent language in [**Her66**] and [**Her67**].

There is a temptation here to maintain that axioms are analytic, but deny that analytic sentences must be true—in other words, to reject the usual sense of 'analytic' in favor of some other sense. This works only to the extent that one is prepared to say exactly what this other sense is. It clearly won't do simply to say that analyticity in the favored sense is just like analyticity in the old sense, but without the implication of truth. There is no clear way to "subtract" the notion of truth away from the notion of truth by virtue of meaning.

So axioms are not analytic sentences, at least not in the usual sense of that term; but then what are they? It won't do simply to say that they are widely and deeply *believed*; if that's all it takes to be an axiom, then Tarski's view is reduced to the claim that there is widespread *error* about truth. Hardly anyone disputes this, and in any case it doesn't constitute a deficiency in our *language*.

Faced with this dilemma, and seeing no way out of it, many philosophers have looked unfavorably on Tarski's view, as I have said. And so I must explain what I mean when I say that 'true' purports to satisfy the disquotational schema, and show how my view avoids this dilemma. I will begin with an explanation which will eventually need an explanation of its own.

My claim is that the fact of our meaning *true* by 'true' consists in our accepting the disquotational schema. This avoids Herzberger's objection, since something can be accepted without being true or even consistent. It also avoids the other horn of the dilemma, since my claim is not simply that we accept an inconsistent schema, but that this acceptance is about all there is to 'true' meaning what it does. It also explains why Liar reasoning seems correct, and why attempts to diagnose that reasoning have proven unsuccessful: those attempts must ultimately treat the disquotational schema as a false theory and propose a replacement, whereas on my account no alternative to the disquotational schema could possibly count as a "correct" account of truth.

At this point an objection presents itself that several authors have made:[6] what exactly do I mean by 'accept' when I say that our competence in using 'true' consists in our acceptance of schema (T)? The reason this is a problem is that apparently we may *reject* that schema and remain competent speakers of English. When faced with an instance of Liar reasoning, for example, an English speaker might choose to reject an instance of (T) rather than accept the inconsistent conclusions it generates. This would not be a very typical reaction to the Liar—just seeing that (T) is what leads to the contradiction requires a level of sophistication about the Liar that relatively few people possess—but it is a possible one, and whatever else one may say about it, it doesn't seem incompatible with remaining a competent English speaker. How, then, can we say that our hypothetical speaker accepts this schema, which she also explicitly rejects?

This argument is particularly forceful if accepting a sentence is the same as being disposed to assent to it. In that case, our speaker is clearly *not* disposed to assent to all the instances of (T), indeed is disposed to dissent from at least one of them, and so clearly she does not accept (T). At least, we must accept this conclusion unless we are willing to maintain that she has a disposition to assent that is being "masked" in this particular situation, in something like the way a sugar cube's disposition to dissolve in water is masked when the water is already supersaturated with sugar. While I think there might be something to such a view, I certainly don't want to insist on it.

For this reason, I don't identify acceptance of a sentence with a disposition to assent to it. Instead, I identify it with a certain kind of *commitment*. The fact of 'true' meaning *true* consists, I claim, in our linguistic conventions committing us to the disquotational schema. This is the same sort of commitment that obtains when a sentence is analytic, but it differs from analyticity in that it is not built into the concept of commitment that what one is committed to is true. It is appropriate to call this commitment "acceptance" because we accept the commitment implicitly,

[6]In different forms, it can be found in [**Par74**] and [**Soa97**].

through our linguistic conventions. I will also call commitments of this sort "linguistic commitments", to distinguish them from other sorts of commitments (e.g., doxastic). I will have more to say later about how such a commitment to a sentence can coexist with a willingness to reject the sentence.

To get a better picture of the sort of commitment I am talking about and how it could arise, let us next consider an example that is artificial but (hopefully) instructive. The idea that one can introduce a new expression into a language by means of a set of stipulations is familiar enough, as is the idea that a truth predicate could be introduced into a language that didn't already have one by adopting some sort of disquotational rule. Now imagine language users whose language lacks a device of disquotation, and who choose to introduce one by means of the following stipulation, by which they intend to define a new predicate G:

(S) The proposition that p is G iff p

Naturally, they do not yet see that there is any problem here. We may assume that all the conditions necessary for the speech act to count as a successful definition have been met.[7] In particular, (S) is intended as an implicit definition of G, and the speakers of this language accept it as such.

Have these speakers succeeded in introducing a meaningful predicate into their language? I would claim that they have, mainly on the grounds that they have evidently conferred a *use* on the predicate G. G functions in their linguistic practice, we may suppose, much like any other predicate. Part of this functioning is that a set of communal *norms* is now in place regarding the use of G: some uses are counted correct, some incorrect. In particular, there is now a commitment to the definition (S). Vague though the slogan "meaning is use" is, it is accurate enough, I would think, to render G meaningful their language.

[7]Or at least all the uncontroversial ones; if consistency of the stipulation is one such necessary condition, then of course it has not been met.

At this point we might recall another well known example of vocabulary introduced by an inconsistent definition, from [**Pri60**]. There A. N. Prior invites us to consider a connective 'tonk', introduced by the rules

$$\frac{A}{A \text{ tonk } B} \qquad \text{and} \qquad \frac{A \text{ tonk } B}{B}$$

Plainly, any sentence whatsoever can be derived from any other sentence when these rules are present. The example was initially devised to cast doubt on the idea that the meanings of connectives are characterized by their inferential roles; and while reactions to the example differed, a consensus emerged to the effect that there are limits to how a new expression can legitimately be introduced via inference rules or axioms.[8] Doesn't my claim that our speakers have introduced a meaningful predicate by (S) ignore the lesson of 'tonk'?

I agree that we *ought* not to try to introduce an expression by means of an inconsistent stipulation, that we would be making a kind of *mistake* if we did so. But that is not the same as saying that that expression would lack meaning. In other words, the proper restrictions (whatever they may be) on introducing an expression are necessary conditions not for the introduced expression to be meaningful, but rather for the introduction of the expression to be unobjectionable. As for 'tonk' itself, I don't think it is meaningful, but the reason is not that the associated rules are jointly inconsistent. I simply can't imagine 'tonk' having a use, that is, I can't imagine it functioning in any language the way connectives like 'and' do, because of how obviously inconsistent the rules are. The schema (S), on the other hand, could easily form the basis for a use that causes trouble only in very special and rare cases.

If all this is right, the predicate G is very much like I am claiming the predicate 'true' is. It is meaningful, as 'true' clearly is. And there is a commitment, due to the conventions of language, to the schema (S), of the same sort that I am claiming we

[8]A popular view, put forth in [**Bel62**], is that the new inference rules must be *conservative* over the preexisting inference rules, i.e., they must not license any new inferences among sentences made of the old vocabulary. Some think this is too restrictive.

have to (T). However, there are some significant differences between the two cases; in particular, doubtless nothing like the act of defining G ever took place for the truth predicate of any natural language. This is a less important difference than it might seem at first, though. G has its meaning not so much because of the act of defining it, but because of the way it is *used*; defining G simply serves to confer that use on it. If we like, we can imagine that our speakers gradually forget all about the act of defining G, but continue to use G in the same way and so continue to mean the same thing by G that they meant initially. In particular, their language would continue to commit them to (S).

Finally, imagine that at some point after the introduction of G has been long forgotten, someone discovers a Liar argument. She would find each step of the argument compelling, since each step is forced on her by her commitments. If she isolated (S) as the source of trouble, she would find it hard to understand how (S) could fail. And if she considered alternative theories of truth, she or her fellow speakers would find them all lacking, and indeed none of them would really count as the correct theory of G-ness. In this way, the situation is again similar to that of truth predicates in natural languages; I will return to this point below.

4. What the Theory Isn't

There are a number of theories of truth that are superficially similar to the inconsistency theory, and I want next to point out the differences, lest my account be confused with something else. It is especially important to distinguish the present theory from those that claim that the Liar is a true contradiction, a statement that is both true and not true.

The motivation for the true contradiction account is simple: the Liar paradox appears to be inevitable, so maybe we should simply accept the inconsistency rather than trying to avoid it or explain it away. This isn't really so different from the motivation for my view, which stresses the apparent unavoidability of the strengthened

Liar. But the conclusion is entirely different. While I think speakers of English are committed to (T), saying so is quite different from *advocating* (T), just as pointing out an inconsistency in a system of beliefs or a conflict in a set of moral obligations is quite different from advocating those beliefs or taking on those obligations. The theory that the Liar is both true and not true is an inconsistent theory, and that is sufficient grounds for rejecting it, at least if the conventional wisdom about inconsistency is correct.[9] If there is any inconsistency in *my* theory, on the other hand, it is not an obvious one.

The best developed version of the true contradiction theory is that of Graham Priest's [**Pri79**] and [**Pri84**]. Priest thinks the Liar is unavoidable and advocates accepting (T) in unmodified form; consequently, he thinks the Liar sentence is both true and not true. He regards the most serious challenge to his view to be the claim that, if we accept a contradiction as true, we will be committed to every sentence whatsoever. To meet this challenge, he develops a nonstandard logic in which a contradiction does not imply every sentence. Now recall from the discussion of Curry's paradox that in the presence of (T), every sentence whatsoever is derivable provided the rules of conditional proof and *modus ponens* are present; thus, Priest's logic cannot contain both rules. In fact, Priest rejects *modus ponens* (at least in unrestricted form).

I won't try to refute Priest. In general, there's no point trying to prove or disprove a logical assumption, since a purported proof would probably have to use the very logic it sought to defend. I will, however, try to undercut some of Priest's motivation: I don't think his considerations provide a strong reason to believe his account of the Liar, unless one is already sympathetic to his account or at least to the view that some contradictions are true.

First, while I agree in general terms with Priest's claim that the Liar is inevitable, the sense in which it is inevitable is one that must be stated with care. A contradiction may certainly be derived from any of various versions of (T); moreover, the task of

[9]Of course, advocates of the true contradiction theory don't accept this conventional wisdom.

finding a variant of (T) that is both free from contradiction and has a reasonable claim to govern the truth predicate as we actually use it, seems hopeless. But in purely formal terms, there are ways around the Liar, e.g., restrictions on (T) that avoid the contradiction. And while these restrictions may be rather implausible, so are the restrictions on logic that Priest resorts to to avoid commitment to all sentences. Thus it is entirely open to us to accept a wholly classical logic, reject true contradictions, and accept a modified (albeit artificial) version of (T); and the cost of doing this, while considerable, is not obviously greater than that of accepting Priest's theory. In other words, Priest regards classical logic as negotiable but (T) as non-negotiable, and provides no clear rationale for such an attitude.

Now I want to be clear: I am not advocating acceptance of a modified, consistent version of (T), except possibly as a revisionary account of truth. Nor am I advocating acceptance of (T) as it stands. To advocate, as a descriptive account, *any* such account of truth would be a mistake on my view, akin to presupposition failure: every use of 'true', on my view, carries with it a commitment to (T), and thus to every sentence of English, and therefore must be rejected.

This brings us to my second objection. As I have already stated, the inconsistency theory and Priest's theory are both motivated by the view that the Liar is unavoidable. The former, to its credit, explains this inevitability without involving itself in the contradictions it describes, thus showing that an inconsist*ent* theory—as opposed to an inconsisten*cy* theory—is not needed to account for the inevitability of the Liar.[10]

Aside from Priest's theory, a view with which mine might be confused is that the truth predicate is an *empty* predicate, i.e., one with an empty extension; or in other words, that nothing at all is true. Such a confusion might be invited by the expression 'inconsistent concept', since one might very well use that term to describe

[10]For a good discussion of Priest's view, see Terrence Parsons' [**Par90**]. His main point is that Priest's account of the Liar is highly parallel to a traditional truth value gap account, to the point where it seems rather arbitrary whether we regard the nonclassical truth values in Priest's truth tables as gaps or gluts.

such concepts as *round square*; the latter is empty precisely because it is inconsistent to say that there are any round squares. The concept of truth is a very different sort of thing. While it is inconsistent to say that there are round squares, there are many perfectly consistent uses of 'round square'—as in 'there are no round squares', for example; but *any* use of 'true' involves a commitment to (T) and must therefore be rejected. This is so even when 'true' is being used to say that nothing at all is true.

One last view that differs from mine is the view that 'true' is *meaningless*. As I have indicated above, I think the truth predicate is perfectly meaningful, but defective in a certain way. Clearly the use of 'true' is governed by standards of correctness, and I would claim that this makes it meaningful. Indeed, it is these very standards that make the Liar a puzzle in the first place.

Of course, to say that 'true' is meaningful is not to deny that it is somehow *defective*. One might even call it *incoherent*. There are many sorts of defectiveness in natural language short of outright meaninglessness, such as reference failure, presupposition failure and category mistakes. Even the word 'nonsense' is ambiguous enough that it might sometimes be used to describe the truth predicate—just not when it is used to indicate literal lack of content.

5. Why Believe the Theory?

Now that the inconsistency theory of truth is on the table, I want to say in more specific terms why I think we should believe it. It is important to make it clear at the outset that "reason to believe" is not the same as "proof". Philosophers have seldom, if ever, managed to *prove* their theories, and I don't expect to be an exception. The inconsistency theory is the best answer I can find to the difficult problem of the Liar, but it isn't the only possible answer. Accordingly, the present section has the more realistic aim of presenting some of the considerations in favor of the theory, and (hopefully) convincing the reader that it deserves to be taken seriously and developed further.

My defense of the theory does make a few assumptions. For one thing, I assume that the last two sections sufficed to make the inconsistency theory at least comprehensible. I also assume that our use of 'true' is governed by *something like* (T). By this I mean that the fact of our meaning what we do by 'true' consists of, or at least involves, linguistic commitment to a rule or rules that coincide with (T) in the ordinary run of cases. This is not an entirely trivial assumption. The claim that meaning *true* by 'true' is nothing other than accepting of (T) is a central component of (at least some versions of) *deflationism* about truth; and deflationism about truth is far from uncontroversial. This assumption is not, however, a substantial assumption *about the Liar*. My assumption says nothing about what the rules governing 'true' say about paradoxical sentences.

Let us call a view that shares these basic assumptions, but holds that 'true' is governed by a consistent rule, a *consistency* theory of truth. My argument for the inconsistency theory is essentially that it better explains the facts than any known consistency theory, and probably explains them better than any as yet undiscovered such theory.

Let's begin by seeing what our practice would be like if the inconsistency theory held. Obviously our practice of ascribing truth would be as it is in most cases: we typically use the truth predicate as a device of semantic ascent and descent in a way that (T) easily facilitates. All theories under discussion are on par in this respect, since they all hold that use of the truth predicate is governed by a rule that agrees with (T) in typical cases. What about atypical cases, specifically cases of Liar reasoning? Here the inconsistency theory also gets the facts right. When presented with an instance of Liar reasoning, a typical reaction is to find each step compelling, to find the conclusion unacceptable, and to be hard pressed to say just where the argument goes wrong. This is exactly what we should expect, assuming the inconsistency theory. Each step is compelling because the reasoning depends on little else than the all-but-analytic (T). We have a hard time discovering what went

wrong because when we try to diagnose a bad argument, we tend to look for either an invalid inference or a factual error—but the Liar argument contains nothing that could appropriately be termed factual error, and we are disinclined to regard its truth inferences as invalid because we are committed to those inferences by the conventions of our language.

How well would a consistency theory explain our practice? To see why it might not explain it very well, let's consider for a moment a rather simpleminded consistency theory, one that holds that the rule governing our use of 'true' is

(T*) If 'S' is OK, then 'S' is true iff S

where 'OK' is a placeholder to be filled in by the theory. What replaces it needs to be strong enough to filter out paradoxical sentences, yet weak enough to let in everyday sentences.

Such a consistency theory would have an easy time explaining our everyday use of the truth predicate but, unlike the inconsistency theory, would have a hard time explaining our reactions to the paradoxes. In particular, if commitment to (T*) is implicit in our use of 'true', why do we persist in trying to apply (T) to Liar sentences when presented with Liar reasoning, rather than simply recognizing that such sentences are not OK? Bear in mind at this point that Liar reasoning remains intuitively compelling even after one recognizes that the relevant instance of (T) is inconsistent; why should this be, if our language commits us only to (T*)?

The only answer consistent with a consistency theory is that even though acceptance of (T*) is implicit in our use of 'true', we have trouble putting (T*) to use in those cases where it disagrees with (T). Perhaps we have a persistent and widespread (but largely inarticulate) folk theory of truth, embodied by (T), which happens to be at odds with our actual commitments but which strongly influences our thinking about the Liar. In other words, maybe we have a major blind spot in our understanding of our own truth concept. Someone who thinks this would certainly have

to explain the disparity between our grasp of (T*)'s consequent, which we have little trouble with, and of (T*)'s antecedent, which we are so appallingly bad at understanding even when we clearly see that (T) by itself is untenable. Perhaps more importantly, we are owed an explanation of why our linguistic intuitions shed no light on (T*)'s antecedent, why they persist in telling us that (T*)'s consequent ought to apply even when we clearly see that it can't. After all, if we are committed to (T*), this is because we have implicitly accepted that commitment in our use of 'true', so the rules governing that use, and hence those commitments, should in principle be available to speakers of English upon reflection in the same way that analytic truths are.

Second, while it's easy enough to see how (T) could have been learned to begin with, it's hard to see how (T*) could have been learned. This point is really two points, one about the evolution of languages and the other about the acquisition of language by individuals. As for language evolution, the primary role the truth predicate plays in language use is to provide a means of agreeing or disagreeing with a proposition (whose exact content may be unknown) and of expressing generalizations about propositions—in short, it serves as a device of disquotation. It's fairly safe to assume that recognition of the semantic paradoxes in a given linguistic community comes only after a truth predicate has long been in place, if ever. Why, then, would the need for a truth predicate result in the acceptance of (T*), whose antecedent plainly serves only as protection from paradox?

Likewise, when a child first learns the meaning of 'true', she does so the same way she learns the meaning of any word: by observing instances of its use, and guessing from them what it means.[11] For a vast majority of children, these observed uses of 'true' involve no paradoxical sentences; and even if they do, they are unlikely to suggest any particular method of avoiding paradox. What those uses of 'true'

[11]At least, this is my understanding of the standard view of language acquisition; for our purposes, it doesn't matter whether this is exactly right.

do involve are what I mentioned in the last paragraph: the truth predicate's use as a device of disquotation. Thus we should expect the child to assume that the meaning of 'true' involves commitment to something like (T), rather than something like (T*). We should therefore expect the inconsistency theory of truth to be a fairly accurate description of most people's ideolects. And since there is no convention among speakers of English (or any language, as far as I am aware) of deferring to Liar specialists about the correct use of 'true', it follows that the inconsistency theory is a fairly accurate description of English (and other languages) as well, excluding perhaps some specialized dialects.[12]

So our simpleminded consistency theory doesn't look very promising. However, most consistency theories are not based on (T*) (at least, not explicitly), and so we need to see how *they* fare. Specifically, we need to reexamine the preceding two objections in the more general context of consistency theories.

Let's begin with the first: how well could a consistency theory explain our reactions to the paradoxes? Here I think the preceding chapters show that their prospects are dim. Insofar as they succumb to the strengthened Liar problem, such theories fail to eliminate all instances of Liar reasoning that are strongly intuitively compelling yet unacceptable, *even after* our intuitions have been informed by the theory. And if such a theory fails to educate our intuitions into consistency, then it must explain why those intuitions remain at odds with the consistent rule(s) that it says we are in fact committed to.

As for the second, that (T*) is not plausibly learnable, some consistency theories will naturally fare better than others. As I indicated in the discussion of Burge's theory, the latter has a particularly hard time explaining both how the concept of truth could have arisen in the first place and how any speaker learns it. Other consistency theories are more plausible on this score. In any case, though, the range

[12]My point is not that (T*) is too *complicated* for us to have learned it as children. Children certainly learn very complicated grammars at an early age, and I see no reason to doubt that they learn complicated concepts as well.

of consistency theories that meet this objection is considerably smaller than the set of all possible consistency theories; this fact reduces the likelihood that an altogether plausible consistency theory will someday be found.

To conclude, then: the persistent failure of past theories to avoid the strengthened Liar problem, combined with the serious obstacles that any consistency theory would have to face, taken together with the ease with which the inconsistency theory handles those obstacles, indicate to me that assuming the basic assumptions of this section, and all else being equal, the inconsistency theory of truth better explains the relevant facts and is altogether likelier than the alternatives. 'All else being equal' naturally conceals a great deal, as there are many potential objections to the inconsistency theory that have not yet been addressed. The next task is to respond to some of those objections.

6. Objections Considered

6.1. Do We Really *Accept* Schema (T). I want to begin by looking more closely at an objection I already touched on. Earlier I noted that the claim that we "accept" (T) must be interpreted cautiously, since there are situations—notably, when faced with Liar arguments—where one might reject some instances of (T). There are a number of objections to views like mine that are based on this fact.

To begin with, the claim that we accept (T) is sometimes taken as the claim that we have a disposition to assent to (T)'s instances. Then the fact that we are sometimes unwilling to assent to a particular instance of (T) is then used to refute this claim. (See [**Par74**] for an argument along these lines.) But the fact of our failing to assent to particular instances of (T) doesn't count directly against the claim that we have a linguistic commitment to (T), since begin committed to a sentence is very different from being disposed to assent to it.

In claiming that what we are disposed to assent to comes apart from what we are committed to, the inconsistency theory says exactly what an account of the concept

of truth should say. *In general*, there is no straightforward relation between meanings and norms on the one hand and dispositions on the other. This point was made forcefully—and famously—in Kripke's [**Kri82**]. To repeat Kripke's main example, suppose Jones means *plus* by '+'; what disposition(s) of Jones could determine this fact? A natural answer, that he is disposed to respond to the question 'What is $x+y$?' with the sum of x and y, simply won't do. There are some cases where Jones has no disposition to respond one way or the other (when x and y are too large, for example), and some where he is actually disposed to respond with some *other* number (since he is disposed to make a mistake).

Admittedly the Liar is somewhat different from the Jones case. Jones's failure to respond with the sum of x and y is a result of his limitations; if he were a bit smarter, or had a bit more time or energy, he would answer correctly in a somewhat wider range of cases. By contrast, failure to endorse an instance of (T) is compatible (on my view at least) with knowing all the facts and thinking clearly. I don't think this counts against the inconsistency theory, however. It simply demonstrates a different way for dispositional facts to be out of sync with normative facts: when commitments come into conflict with each other, not all of them will be honored, even if they are all recognized.

In fact, we can find clear examples of just this sort of thing for other kinds of normative facts. People's beliefs often commit them to mutually inconsistent propositions, yet such people are seldom prepared to assert both propositions in the same breath. Thus, beliefs provide examples of commitment to p without willingness to assent to p. In fact, the case of belief is really rather similar to what I am claiming about the concept of truth: the reason for the lack of assent is simply the recognition of inconsistency among one's commitments.

Someone could object that once one recognizes mutually inconsistent propositions p and q among one's beliefs, one then ceases to believe both p and q, or at least to fully belief them. However, the person in my example might believe p and q because

they are consequences of some inconsistent belief r, and what's more, she might not have tracked down r as the source of p and q. In this case, whatever attitude she might adopt toward p and q, she remains committed to them as long as she continues to believe r; so my example remains one of commitment without willingness to assent.

Another area in which dispositions and norms can come apart is that of inconsistent rules. The rules of a game may turn out to be inconsistent in the sense that a situation can arise in an otherwise correctly played game in which no move is legal, yet some move is required.[13] For definiteness, let's imagine an inconsistent variant of chess. Let chess* be the game with all the rules of chess, plus one additional rule: when one of my pawns is in position to capture your queen, when it's my turn, and when I am not currently in check, I must capture your queen with one of my pawns. (By 'in position to capture your queen' I simply mean that one of my pawns is diagonally adjacent to your queen and is closer to my side of the board. 'You' and 'I', of course, stand for any two players.) The game is inconsistent, since while I might not be currently in check, it might happen that I have a pawn in position to capture you queen but that in so doing I would put myself in check. In that situation no move is legal (since chess* has all the rules of chess, including the rule that one may not put oneself in check). Still, it's easy to imagine that this inconsistency would go unnoticed for a while, and that some people might actually play chess*.

Now to be someone who plays a given game is not the same as behaving in accordance with its rules, since people cheat, make mistakes, etc.: it is for its rules to determine how one *ought* to behave. And in the case of an inconsistent game, it is sometimes impossible to behave in accordance with the rules, even while one obligated to do so. So here again we see dispositions and norms diverging; and again, the situation is parallel in certain respects to that of inconsistent linguistic commitments.[14]

[13] I'm told that the rules of baseball have been shown to be inconsistent in this sense, though I have not seen the proof myself or found a source.

[14] This is especially so insofar as the familiar comparison of languages with games is taken seriously.

So much for dispositions. A somewhat different but related objection in [**Soa97**] simply takes the fact that a competent speaker of English may reject some instances of (T) to show that acceptance of (T) is not a necessary condition for being a competent speaker of English. This objection may be sound if 'accept' is being used in its ordinary sense, since arguably one cannot both accept and reject something. But as I am using (or perhaps misusing) the term, to accept a sentence or schema is just to be linguistically committed to it. And it is certainly possible to be committed, linguistically or otherwise, to something one rejects; in rejecting (T) one may be speaking English incorrectly, but one is still speaking English.

Another objection points out that in practice, we do a pretty good job of avoiding paradox in our use of 'true', in the sense that we manage to use it without getting into too much trouble. To borrow an example from McGee, you would be sorely disappointed if you tried to argue your way out of a speeding ticket with an argument that began "consider A, your honor, where A is the sentence, 'If A is true, then I wasn't speeding'...." Doesn't this show that the rule we accept is not (T) but rather something consistent?

I claim it doesn't. It is more likely that our ability to use the truth predicate without getting into too much trouble is a special instance of a more *general* ability not to let contradictions get us into too much trouble. In the example just given, you would fare no better if, discovering that the judge had previously asserted mutually inconsistent propositions p and q, you argued in a standard way from p and q to 'I wasn't speeding.' When one recognizes a contradiction in one's system of beliefs, one doesn't go on to derive consequences from that contradiction; one rejects the offending belief, or at least refuses to use the belief in argument. And when we "handle" the Liar in practice, I think we are doing essentially the same thing: relying on a general ability not to let inconsistency infect our belief system.

Actually, a little bit more needs to be said here since some instances of "avoiding the paradoxes" in practice do not involve any actual inconsistency. Consider the

following instance of Curry's paradox:

(3) If (3) is true, then Kaczynski is the unabomber.

Using (T) we can conclude that Kaczynski is the unabomber; and just about everyone who could follow our reasoning would find it unconvincing. In this case, however, the relevant instance of (T) is consistent, and not only is the conclusion consistent, but most people would agree with it. So no general facility for managing inconsistent belief could explain what's going on here.

What we see here is rather a general disposition for detecting and distrusting specious argument. The argument that Kaczynski is the unabomber is suspect because it could be adapted to establish anything at all; for this reason, I suspect, people distrust it even if they can't see anything wrong with it. Again, however, we don't have any *specific* strategy for avoiding semantic paradox.

Finally, this discussion raises an important question: just what does the fact of our being committed to (T) consist in? This is a special case of a more general question: how do the *natural* facts about us determine the *semantic* and *normative* facts? This question has received a great deal of attention,[15] and so far it has resisted solution. While most philosophers would agree that what we mean by our words is determined by how we use them, the relation between these two things is not well understood. So I don't know what natural fact our commitment to (T) consists in; nor does anyone else know what facts make 'true' mean what it does. So when asked what it is for us to be linguistically (or otherwise) committed to (T), I can only respond that I don't know, and that it isn't my job to say.

But while perhaps no one is in a position to say what it is for a given claim about meaning to hold, we do often present evidence for and against such claims. For example, while we may not be able to explain what it is for Jones to mean *plus* by '+', we can certainly have evidence that he does: his account of the procedure he uses

[15]Much of it stemming from [**Kri82**]

for adding, the fact that he usually does give the sum of x and y when asked what '$x + y$' is (at least for small x and y), perhaps even the kind of mistakes he tends to make. The arguments of the last section should be understood in this spirit.

6.2. Deflationary Truth. In claiming that accepting (T) is all there is to meaning *true* by 'true', I am endorsing what has come to be known as a *deflationary* conception of truth. Deflationism has historically been a minority position among philosophers (though perhaps the same could be said for any account of truth). I won't try to defend deflationism here, partly because I have little to add to Horwich's defense of it in [**Hor90**]; what I will do is show that some of the reasons for being unhappy with deflationism don't apply to the present version.

First, recall that there are two ways of viewing truth: as a feature of sentences, and as a feature of propositions. This in turn leads to two different versions of deflationism, a sentential and a propositional version. The sentential version takes truth to apply primarily to sentences, and holds that all there is to truth is the schema

($\mathrm{T}_{\mathrm{Sent}}$) '$S$' is true iff S.

A well known objection to this view is that it has a hard time accounting for the fact that whether a sentence is true depends on what it *means*, which in turn depends on contingent, empirical facts about language users. If ($\mathrm{T}_{\mathrm{Sent}}$) is analytic, as presumably it must be for sentential deflationism, then the fact that, for example, 'Snow is white' is true just in case snow is white, is also analytic. But surely this fact is *not* analytic, since it would not have obtained if 'Snow is white' had meant that snow is black, for example.[16]

The other version of deflationism takes truth to apply primarily to propositions, and only derivatively to sentences: a true sentence on this view is one that expresses

[16]I realize there is more to be said here; I am only trying to summarize a familiar argument, not make a conclusive case.

a true proposition. It holds that all there is to truth is the schema

(T_{Prop}) The proposition that p is true iff p.

This version of deflationism has no trouble accounting for the dependence of sentential truth on meaning, since it allows that what proposition a given sentence expresses depends on contingent, empirical facts about speakers. In fact, for propositional deflationism, sentential truth can be as "robust" as you like. The version of deflationism I assume is propositional; it therefore avoids the above problem.

Second, some versions of deflationism take the rather extreme view that truth is "not a property,"[17] while some others make the more moderate claim that while truth is a property, it is not a "substantial" property. The version of deflationism that I assume certainly does not make the former assumption. Even if all there is to truth is something like (T), 'true' still applies to some propositions and not to others, and there seems to be no obstacle to saying that those propositions it applies to—the true ones—have the property of being true. As to the second claim, I have to admit that I really don't know what people mean when they argue over whether truth is a "substantial" property. Maybe truth's lack of substance follows from the claim that (T) is all there is to truth, and maybe it doesn't; I simply don't want to make any *additional* assumption regarding this issue.

Now in a way the inconsistency theory *does* seem to imply that there is no such property as truth (although of course there is a *concept* of truth). After all, any such property would have to satisfy (T), which is impossible. Moreover, assuming the inconsistency theory, one should not say, as I just did, that some propositions are true and others are not, since any such use of 'true' carries with it commitment to (T). However, the version of deflationism I assume in arguing for the inconsistency theory—that acceptance of *something like* (T) is what makes us mean what we do by 'true'—is neutral on the question of consistency, and is therefore compatible with

[17]The most developed version of this position is that of [**GCB75**]

the idea that 'true' picks out a property. So while I may be arguing that there is no such property as truth, I don't assume this at the outset. In any case, I think the truth predicate *purports* to pick out a property, whereas advocates of a more extreme deflationism deny even this.

Next, deflationism sometimes takes the form of a claim about the nature of truth ascriptions considered as *speech acts*.[18] According to a deflationism of this stripe, an apparently assertive utterance of the form 'p is true' is not being used to *say* anything at all, but rather to *do* something, namely, to endorse or agree with some assertion. However, deflationism need not take this form, and the present version certainly does not.

Finally, the Liar itself is sometimes used as grounds for criticizing deflationism. Namely, if deflationism entails that (T) is all there is to truth, and if (T) is inconsistent, isn't deflationism inconsistent? Horwich considers this objection, and replies by claiming that it is not (T) itself but some suitable restriction of (T) that characterizes truth, and that it is not up to him to say what that restriction should be.[19]

Horwich is forced to say this (or at least to say something along these lines) by the fact that believing his theory requires endorsing (T), or rather some suitable restriction of (T). So it is with most versions of deflationism. The inconsistency theory provides a way around this. It claims that in a certain sense, (T) is all there is to truth, and in that respect it is deflationary; but rather than actually endorsing (T), it claims that acceptance of (T) is all there is to meaning what we do by 'true'. Thus, the inconsistency theory saves deflationism from a longstanding worry.

[18]See [**Str50**], for example.

[19]Mark Kalderon has raised an interesting objection to Horwich's strategy for handling the Liar. For Horwich, truth is characterized by some consistent set of instances of (T_{Prop}), and the instances in this set are analytic, while those not in the set are not. But as Kripke pointed out in [**Kri75**], some paradoxical sentences are paradoxical only contingently: they generate paradox if, but only if, the facts turn out to be a certain way. (The Nixon-Dean sentences are of just this sort, as is the sentence 'If Clinton was re-elected in 1996, then this entire sentence is untrue'.) In fact, many of the sentences we routinely use turn out to be risky, in that they will be paradoxical if the facts turn out to be particularly unfavorable. So the restricted set of T-biconditionals must include some that involve contingently paradoxical propositions, lest (T) be too restrictive. But then those instances can hardly be analytic.

6.3. Truth and Meaning. A popular tradition in the philosophy of language holds that a sentence's meaning is simply its truth conditions, and that a word's meaning is the contribution it makes to the truth conditions to the sentences in which it occurs. Such views tend to come in two flavors. First, there is the idea, expressed in Donald Davidson's [**Dav67**], that to give a theory of meaning for a given language it is enough to give a finitely axiomatizible theory that has as first-order consequences all the T-sentences for that langugage. Second, there is a family of views that identify the meaning of a sentence with the set of "circumstances" in which it is true, where a circumstance may be a possible world, or something slightly different, depending on the theory. Those who hold such views have (at least) two reasons to be unhappy with the inconsistency theory.

First, if meanings are truth conditions in the *ordinary* sense of 'true', then any incoherence in the concept of truth automatically becomes an incoherence in the concept of meaning. While rejecting the concept of meaning may be an option for some philosophers, it is not an option for me, since I rely rather heavily on that notion: the inconsistency theory is really a theory about what 'true' means. Naturally, then, I must reject the view that meanings are truth conditions in the ordinary sense of truth.

Actually, depending on the precise form that the truth conditional theory of meaning takes, it may already be incompatible with the deflationary view of truth I have been assuming. Namely, deflationism is incompatible with any view that regards the concept of truth as explanatorily prior to that of meaning. To see this, recall that on the above version of deflationism, 'true' applies (or purports to apply) primarily to propositions, and a true sentence is simply a sentence that expresses a true proposition. The notion of a true *sentence* therefore cannot be used to explain the notion of meaning, since it presupposes the notion of a sentence expressing a proposition; and if the notion of a proposition is not actually identical to that of a meaning, it is

close enough that an account of meaning would be rather trivial if it presupposed the notion of a sentence expressing a proposition.

Of course, deflationism about truth is controversial, and in any case there may be variations on the truth conditional theory of meaning that the above argument doesn't address. My goal here is not to argue against the theory so much as to (a) acknowledge that it is probably incompatible with the inconsistency theory, and (b) point out that it is by no means universally accepted.

The second source of potential incompatibility is that a truth conditional account of meaning seems to preclude the very idea of an inconsistent concept. It just isn't at all clear how the concept of linguistic commitment could be explained in terms of truth conditions, if the existence of a linguistic commitment to a sentence is not to imply that that sentence is true. If we take the sentences a language commits its speakers to to be those that are true in that language under all possible circumstances, for example, then clearly our linguistic commitments must be true, and the inconsistency theory is rendered inconsistent.

I think this idea is behind many extant objections to the inconsistency theory. It is certainly behind the argument of [**Her67**]. There Herzberger argues that there are no inconsistent languages, but he takes an inconsistent language L to be one where there is a set S of sentences of L and a logically possible condition c such that (a) every sentence of S is true in L in the condition c, and (b) S is inconsistent. It's no great surprise that an inconsistent language in *this* sense cannot exist, since if one did there would be at least one sentence in S that was both true and not true in c. This characterization of inconsistent language is apparently the best Herzberger can offer given the general assumption that meanings are truth conditions; I suspect it is about as good as anyone can do.

In the same article, Herzberger makes a remark that bears on something we said earlier. If an inconsistency conception of truth must be rejected, then the inferences people make in Liar reasoning are factual errors, however seductive they might be.

Herzberger writes that "Such mistakes provide... a particularly incisive refutation of the philosophical view that the facts of language are transparent to native intuition." [**Her67**, p. 35] In more than one place above, I assumed that our linguistic commitments are in principle accessible to us via our linguistic intuitions. Someone who thinks that meanings are truth conditions, or more generally who thinks the inconsistency theory is incoherent, will probably be forced to agree with Herzberger on this point.

I think my assumption is more natural than Herzberger's, at least on the matter of linguistic commitments. It is also more in accord with the other examples of non-linguistic commitment mentioned above, namely beliefs and games. While the naturalness of a view is hardly the same as a decisive proof, it is certainly a consideration in a view's favor. It would be one thing if we really were forced to reject the accessibility to intuition of linguistic commitments; but as far as I can see, we are forced to do so only if we reject the inconsistency theory. It therefore seems reasonable to assume the accessibility of linguistic commitments when arguing for that theory.

6.4. *Tu Quoque?* Finally, we must consider whether the inconsistency theory is self-defeating in much the way the views criticized in Chapters 3 and 4 are. While I claim we must reject the concept of truth, I do sometimes use the word 'true', along with its synonyms, in this chapter. I also use others, like 'inconsistent', which might be thought to presuppose the notion of truth. If these uses of true turn out to be indispensible, or if the notion of inconsistency turns out to presuppose that of truth, then the inconsistency theory, or at any rate my defense of it, is fatally flawed.

So let's consider whether my presentation of the inconsistency theory really does make essential use of a truth concept. Some of the above uses of the truth predicate can naturally be dismissed as rhetoric, or as window dressing of some similar sort. Some others occur inside scare quotes, or within the scope of propositional attitude

operators, as when I say that a certain sentence or inference "seems correct" to most speakers. I assume that there is no serious problem with any of these.

In some other cases, 'true' is simply being used as a device of disquotation. Now this alone does not exhonorate those uses, since it is precisely as a device of disquotation that the truth predicate is problematic. However, most (though certainly not all) of the disquotational work the truth predicate does could be done by some replacement concept governed by some consistent approximation to (T). (I discuss one such replacement concept in the next chapter.) Thus, we may at least hope that the above disquotational uses of 'true' will prove dispensible.

How realistic is this hope? It is impossible to say for sure in the absence of a complete reconstruction of the foregoing sections, which might even need to include formalization. But I think there is reason to be optimistic. In practice, the overwhelming majority of uses of 'true' are very far from the paradoxes and would surely be unaffected by the sort of revision in the truth concept that we are considering. Indeed, many of these uses are what we might call "expandable." Consider a collection of uses of the truth predicate, all of which ascribe truth to a proposition that is actually specified, or that deny that some specified proposition is true. If we replace each sentence 'x is true' by a sentence expressing the proposition that 'x' refers to, we will certainly have preserved the intent behind the original sentences. And if repeated application of this process of "expanding" truth ascriptions eventually results in a collection of sentences in which the truth predicate does not occur at all, then I think it is safe to say that the original sentences can be replaced by sentences that are unproblematic but which serve the same communicative intent. Now I suspect that all, or nearly all, of my casual uses of the truth predicate are expandible in this way, and thus not seriously problematic, albeit somewhat sloppy. But I admit that this is handwaiving; the reader will simply have to assess the prospects for eliminating uses of 'true' from the foregoing for him- or herself.

A potentially more serious problem involves the word 'inconsistent'. The notion of inconsistency is sometimes explained in terms of truth: an inconsistent sentence is one that could not be true, or that is not true in any interpretation of its nonlogical vocabulary. Obviously I need a notion of inconsistency that is different from this.

One sense of 'inconsistent' that is available to me is a syntactic one. When I say that the rules we accept as governing the truth predicate are inconsistent, what I mean is that they license the derivation of every sentence. So what the inconsistency theory of truth really states is that the linguistic commitments generated by the conventions governing the truth predicate in turn generate a commitment to every sentence of our language.

Now one might complain at this point that this account of inconsistency leaves out any explanation of why inconsistency is objectionable. If inconsistency is defined in terms of truth, then the reason is clear: it is wrong to assert something inconsistent because doing so would involve asserting something untrue, for example. But this explanation is not available to me. It is still open to me, however, to point out that there are many sentences that we would like very much not to assert, or that we would strongly prefer not to be committed to; this seems to explain what's bad about having inconsistent commitments.

There is also a somewhat different use of 'inconsistent' that I make above, when I say that a given set of rules is inconsistent. In a similar vein, I sometimes describe our linguistic commitments as coming into conflict with each other. Such talk should be understood as follows. The commitments generated by (T) are positive, in the sense that they are all commitments to a sentence. But other linguistic commitments are negative, in that they are prohibitions: they say that one ought not to assert a given sentence, or a given set of sentences. For example, the linguistic conventions governing negation prohibit the joint assertion of a sentence and its negation. Let us say that two commitments come into conflict when they cannot be jointly fulfilled.

Then a commitment to each and every sentence of a language would certainly come into conflict with the prohibition just mentioned.

I also said in at least one place that the use of 'true' is governed by "standards of correctness." This really ought to be understood as meaning that the conventions of language license or prohibit certain assertions or inferences—something that doesn't appear to presuppose the concept of truth. 'Correct' can be understood as meaning roughly 'in accordance with the rules of language'. Now assuming the inconsistency theory, there are situations where it is impossible to act in accordance with the rules governing the truth predicate. But this doesn't undermine the claim that the use of the truth predicate is governed by standards of correctness; it simply means that there are situations in which the rules of language cannot all be complied with. (If all this talk of language's rules makes you uneasy, ignore the last three sentences and focus on the second sentence of this paragraph.) In general, when the reader detects terms like 'correct' playing a nontrivial role in the preceding sections, he should consider whether they might be construed in a normative, rather than an alethic, sense.

CHAPTER 6

Conclusion

Having now defended the inconsistency theory, I want to end by considering some of its upshot. My remarks will be brief; having finished my main task, I now want only to sketch possible directions for further thinking on these matters.

Let me begin by saying that I am not agitating for any actual linguistic reform, nor am I recommending that philosophers refrain from using the truth predicate in their ordinary use of language. There is a sense in which we must reject the truth predicate, but it is not an all-things-considered sense of 'must'. While those who use 'true' in its ordinary sense are certainly making a kind of mistake, there are worse things than making mistakes. When practical considerations are allowed into consideration, there's no contest: for most purposes, it's better by far to continue using our inconsistent truth predicate than to abstain from using it, or even to replace it with a consistent one.

Not all uses of 'true' are everyday uses, however, and for certain specialized purposes it may be very sensible, all things considered, to do some conceptual revision. One such specialized purpose is mathematical logic, and here the actual history of logic is instructive. The notion of truth that Tarski defined for formalized languages is not the ordinary notion of truth, and it is not inconsistent. Tarski effectively replaced the ordinary notion of truth, which I claim is inconsistent, with a different, consistent notion. And it is clear that the standards of rigor that prevail in mathematics would make mathematical use of an inconsistent concept inappropriate. The discovery of a consistent truthlike concept in logic has also, of course, proven enormously fruitful.

So not only *should* we replace our inconsistent truth predicate with a consistent substitute in certain specialized areas of discourse; in at least one of them, we already have.

This naturally raises the question whether philosophy is more like logic or more like mathematics in terms of the need to revise the truth predicate. Here I think the answer depends on the area of philosophy, and maybe even on the particular philosophical discussion. For the most part, I think the best course is to do what we do in everyday speech and continue to use an inconsistent truth predicate. Granted, abandoning the concept of truth altogether or replacing it with a consistent approximation would make a given philosophical discourse more accurate, in a certain sense; but accuracy is not a virtue that trumps all others, even in philosophy. It is perfectly common in philosophy to forego some accuracy when that accuracy would be unilluminating and distracting—to see this, just consider how often the words 'I will ignore…', 'For simplicity I shall assume…', and the like occur in philosophical writing.

For a few areas of philosophical inquiry, however, finding a consistent replacement for the truth predicate may be a worthwhile goal. An especially pertinent example is the project carried out in this essay. In explaining and defending the inconsistency theory of truth, I did sometimes use the word 'true', or at least synonyms like 'holds'; and according to the very theory I was advancing, this is something that strictly speaking I shoudn't do. If this use of the truth predicate turns out to be unavoidable, then this is a very serious problem for the inconsistency theory. Now I have already given some general reasons to think that these uses will turn out to be avoidable; but the better idea we have about how our truth concept might be replaced by a consistent approximation, the better position we will be in to judge this sort of thing.

Moreover, there is an obvious *theoretical* interest in how the naive truth concept might be replaced by a serviceable consistent approximation. According to the inconsistency theory, our conceptual scheme suffers from a certain deficiency; and in

general it is an important task of philosophy to show how our conceptual scheme might be improved.

I won't make any definite proposal for fixing up the truth predicate, but I would like to indicate in general terms how this could be done. What a consistent approximation to the truth concept should be like depends, of course, on what we propose to do with it, and one of the main things (I would say *the* main thing) we do with the truth predicate is to use it as a device of disquotation. Here it should be fairly clear that very nearly all the work the truth predicate does in practice could be accomplished by a consistent replacement; at the same time, the negative results in this essay strongly suggest that no such concept will be able to do *quite* what we would like from a device of disquotation.

By way of illustration, let's look at one possible replacement concept, which we might call "grounded-truth." Intuitively we may think of a grounded-true sentence as one that is both grounded and true, though of course this cannot be the official definition of the term. Instead, we ought to define the term by adopting an appropriate disquotational rule (about which I will have more to say in a moment). Grounded-truth is a consistent notion, and in the normal run of cases it behaves just as we expect truth to behave. For many purposes, it might admirably serve the need for a device of disquotation. But there is a limit to how well it can serve. The sentence

(1) (1) is not grounded-true

is naturally not grounded-true; but the sentence '(1) is not grounded-true', which says this, is also not grounded-true. This simply represents the limits of grounded-truth as a device of disquotation. For some purposes those limits would be irrelevant, while for others they would be very relevant.

Now groundedness is a technical notion, applying to sentences of a formalized language; if the foregoing approach were to be applied in a natural language, it would have to be modified. One reason, of course, is that the natural language features that

formalized languages ignore would finally have to be acknowledged. But a different reason is that the construction of Kripke's least fixed point for a formalized language takes place in an essentially richer metalanguage. If we are to define a notion of grounded-truth for our own language, we will have to either (1) define it only for a proper fragment of our language, (2) introduce some new concept into our language but then restrict the notion of grounded-truth to our original language, or (3) find a way to define a notion of grounded-truth without recourse to an essentially richer metalanguage.

This last option is more feasible than it might at first seem. For the sake of discussion, let us assume that English (or the portion of English we are interested in) has been formalized as a classical first-order language $\mathcal{L}$ with finite vocabulary. Let us also assume that $\mathcal{L}$ has a unary function symbol $\dot{\neg}$, which we will think of as expressing the operation of negation.[1] Now let G be a unary predicate not in $\mathcal{L}$ and let $\mathcal{L}^+ = \mathcal{L} \cup \{G\}$. We will first define sentences φ^+ and φ^- for every sentence φ of $\mathcal{L}^+$, as follows:

$$\varphi^+ = \varphi, \qquad \varphi \text{ an atomic sentence of } \mathcal{L}^+$$

$$\varphi^- = \neg\varphi, \qquad \varphi \text{ an atomic sentence of } \mathcal{L}$$

$$G(t)^- = G(\dot{\neg}(t))$$

$$(\neg\varphi)^+ = \varphi^- \qquad\qquad (\neg\varphi)^- = \varphi^+$$

$$(\varphi \vee \psi)^+ = \varphi^+ \vee \psi^+ \qquad\qquad (\varphi \vee \psi)^- = \varphi^- \wedge \psi^-$$

$$(\exists x\, \varphi)^+ = \exists x\, \varphi^+ \qquad\qquad (\exists x\, \varphi)^- = \forall x\, \varphi^-$$

Basically, φ^+ and φ^- are equivalents of φ and $\neg\varphi$ in which G only occurs positively (i.e., within the scope of an even number of negation signs), where we regard $G(\text{`}\neg\varphi\text{'})$

[1] In what follows we will assume that the sentence $\dot{\neg}(\text{`}\varphi\text{'}) = \text{`}\neg\varphi\text{'}$ is provable in the theory T and holds in the model $\mathfrak{M}$, for each sentence φ of $\mathcal{L}^+$, which we will define presently.

and $\neg G(\text{`}\varphi\text{'})$ as equivalent. The $^+$ and $^-$ operations have the following useful properties.

LEMMA 1. *If φ is a sentence of $\mathcal{L}^+$, then relative to the strong Kleene scheme, (a) for any partial model $\mathfrak{M}$ for $\mathcal{L}^+$, $\mathfrak{M} \models \varphi$ iff $\mathfrak{M} \models \varphi^+$ and $\mathfrak{M} \dashv\models \varphi$ iff $\mathfrak{M} \models \varphi^-$; and (b) if $(\mathfrak{N}, (E, A))$ is a partial model for $\mathcal{L}^+$, then $(\mathfrak{N}, (E, A)) \models \varphi^+$ iff $(\mathfrak{N}, E) \models \varphi^+$ and $(\mathfrak{N}, (E, A)) \models \varphi^-$ iff $(\mathfrak{N}, E) \models \varphi^-$.*

PROOF. Both are proven by an easy induction on φ's complexity. In the case of (b), the crucial fact is that since G occurs only positively in φ^+ and φ^-, its antiextension is irrelevant. □

Now consider the classical first-order theory with the following axioms (for all φ of $\mathcal{L}^+$):

(1) $\forall x \neg[G(x) \wedge G(\dot{\neg}(x))]$;

(2) $G(\text{`}\neg\neg\varphi\text{'}) \leftrightarrow G(\text{`}\varphi\text{'})$;

(3) $G(\text{`}\varphi\text{'}) \leftrightarrow \varphi^+$.

Call this theory F. (The theory F, and the following results, are due to Feferman; see [**Fef82**].) Obviously F is recursively axiomatizible. F is intimately related to Kripke's construction, as the following shows.

THEOREM 1. *Let $\mathfrak{M}$ be any classical model for $\mathcal{L}$; then relative to the strong Kleene scheme, F holds in an expansion $(\mathfrak{M}, E)$ of $\mathfrak{M}$ to $\mathcal{L}^+$ just in case G is a truth predicate for $(\mathfrak{M}, (E, A))$ under the strong Kleene scheme, for some set A.*

PROOF. ($\Longleftarrow$) Assume G is a truth predicate for $(\mathfrak{M}, (E, A))$. Clearly axioms 1 and 2 hold in $(\mathfrak{M}, E)$. To see that axiom 3 holds, notice first that $(\mathfrak{M}, E) \models G(\text{`}\varphi\text{'})$ iff $\varphi \in E$ iff $(\mathfrak{M}, (E, A)) \models \varphi$, and by Lemma 1, $(\mathfrak{M}, (E, A)) \models \varphi$ iff $(\mathfrak{M}, (E, A)) \models \varphi^+$. Finally, $(\mathfrak{M}, (E, A)) \models \varphi^+$ iff $(\mathfrak{M}, E) \models \varphi^+$, again by Lemma 1.

($\Longrightarrow$) Assume F holds in $(\mathfrak{M}, E)$ and let $A = \{\varphi : \neg\varphi \in E\}$. Since axiom 1 holds in $(\mathfrak{M}, E)$, E and A are disjoint and $(\mathfrak{M}, (E, A))$ is a partial model. To show

that G is a truth predicate in $(\mathfrak{M}, (E, A))$, we first show that $(\mathfrak{M}, (E, A)) \models \varphi$ iff $(\mathfrak{M}, (E, A)) \models G(\text{`}\varphi\text{'})$ for all $\varphi$: $(\mathfrak{M}, (E, A)) \models \varphi$ iff $(\mathfrak{M}, (E, A)) \models \varphi^+$ iff $(\mathfrak{M}, E) \models \varphi^+$, and since $(\mathfrak{M}, E)$ satisfies axiom 3, $(\mathfrak{M}, E) \models \varphi^+$ iff $(\mathfrak{M}, E) \models G(\text{`}\varphi\text{'})$.

It remains only to show that $(\mathfrak{M}, (E, A)) \models \neg\varphi$ iff $(\mathfrak{M}, (E, A)) \models \neg G(\text{`}\varphi\text{'})$. By what we just proved, $(\mathfrak{M}, (E, A)) \models \neg\varphi$ iff $(\mathfrak{M}, (E, A)) \models G(\text{`}\neg\varphi\text{'})$, and we know that $(\mathfrak{M}, (E, A)) \models G(\text{`}\neg\varphi\text{'})$ iff $\neg\varphi \in E$; but by the definition of $A$, this holds iff $\varphi \in A$, iff $(\mathfrak{M}, (E, A)) \models \neg G(\text{`}\varphi\text{'})$. □

Now this is not yet directly relevant to the project of introducing a grounded-truth predicate into a natural language, since we are still using an essentially richer metalanguage to talk about $\mathfrak{M}$. However, we can use the model-theoretic result just proven to obtain a relevant proof-theoretic result. A theory $\mathsf{T}' \supseteq \mathsf{T}$ is said to be a *conservative extension* of T if every theorem of T' that belongs to the language of T is also a theorem of T.

THEOREM 2. *If* T *is a theory in the language* $\mathcal{L}$*, then* $\mathsf{T} \cup \mathsf{F}$ *is a conservative extension of* T.

PROOF. Suppose some sentence φ of $\mathcal{L}$ is *not* a theorem of T; we will show that it is not a theorem of $\mathsf{T} \cup \mathsf{F}$ either. Since T does not prove φ, $\mathsf{T} \cup \{\neg\varphi\}$ is consistent and therefore has a model $\mathfrak{M}$. By Theorem 1, $\mathfrak{M}$ has an expansion $(\mathfrak{M}, E)$ to $\mathcal{L}^+$ that satisfies F; thus $\mathsf{T} \cup \mathsf{F} \cup \{\neg\varphi\}$ holds in $(\mathfrak{M}, E)$, and so φ is not a theorem of $\mathsf{T} \cup \mathsf{F}$. □

In particular, F is consistent; but more generally, adopting F is safe, in that doing so will not permit us to prove any nonsemantic facts (i.e., non-G facts) that we couldn't prove already. F's consistency may be somewhat surprising, since axiom 3 resembles schema (T). What happens, for example, when φ says $\neg G(\text{`}\varphi\text{'})$? In this case, since $(\neg G(\text{`}\varphi\text{'}))^+$ is simply the sentence $G(\neg(\text{`}\varphi\text{'}))$, which is provably equivalent to $G(\text{`}\neg\varphi\text{'})$ (see note 1), the relevant instance of axiom 3 is equivalent to $G(\text{`}\varphi\text{'}) \leftrightarrow G(\text{`}\neg\varphi\text{'})$, which is perfectly consistent.

Adopting F is not the ideal way to introduce a grounded-truth predicate into a natural language, however. Although F's underlying logic is classical, the extension of G is provably not closed under classical consequence. To see this, let λ be a sentence that is provably (in T) equivalent to $\neg G(\text{`}\lambda\text{'})$; by Theorem 1, the sentence $\lambda \vee \neg\lambda$ does not belong to the extension of $G$ in any model that satisfies $\mathsf{T} \cup \mathsf{F}$, so $\neg G(\text{`}\lambda \vee \neg\lambda\text{'})$ is a theorem of $\mathsf{T} \cup \mathsf{F}$. So accepting F would amount to letting classical logic govern one's assertions but holding that grounded-truth is governed by strong Kleene logic. This is an infelicity rather than a fatal flaw, since we are clearly free to define grounded-truth any way we like; but we would still like to do better. F was meant merely to illustrate how we could introduce a replacement for the concept of truth while avoiding appeal to a richer metalanguage; it was not meant to be the last word on how we might do so.[2]

We have so far been focusing on the truth predicate's use as a device of disquotation; and beyond that use, there is precious little agreement about the legitimate philosophical uses that a truth predicate might serve. Since I accept a deflationary conception of truth, I tend to think that it has few legitimate uses aside from its disquotational use. Be that as it may, exploring the implications of the inconsistency theory for other philosophical uses of the truth predicate is beyond the scope of this essay.

Finally, I want to consider the inconsistency theory's import for the philosophy of language; and here I can do no more than point in the general direction of further theorizing. I have argued that one particular concept is inconsistent; but if a natural languages possesses one inconsistent concept, it almost certainly possesses others as well.

[2]It is also questionable whether the concept introduced by adopting F would really be one of *grounded*-truth, since any fixed point can serve as the basis for a model of F, not just the least fixed point. Thus, for example, F does not prove that the truth-teller is not G. This doesn't affect the serviceability of the concept, just the choice of terminology.

More generally, natural languages very likely possess words that express what we might call *non-conservative* concepts. Let us say that a word w expresses a non-conservative concept, relative to a given language L, if the rules governing w's use in L license assertions or inferences within the w-free part of L that are not licensed by the other rules. Inconsistency is simply a limiting case of non-conservativeness, where every possible assertion is licensed.[3]

To give just one example, Robert Brandom has made a case for treating pejorative terms as non-conservative.[4] Consider 'pig', for example, used as a derogatory term for the police. Brandom would say that if one accepts this use of 'pig' as part of one's language, then one is committed to the inference from 'x is a police officer' to 'x is a pig' and from 'x is a pig' to 'x is brutal and corrupt' (or something like that). Taken together these yield a commitment to the inference from 'x is a police officer' to 'x is brutal and corrupt', an inference not licensed by the conventions governing 'police officer' or 'brutal' or 'corrupt'. If this is right, then 'pig' expresses a non-conservative concept, and this concept must be rejected if one wants to reject the latter inference.

A familiar complaint about Tarski's conception of an inconsistent language is that it confuses languages with theories. It is theories, the complaint goes, and not languages, that are consistent or inconsistent, and it is theories, not languages, that make claims. But if what I have been saying is right, languages have more in common with theories than this complaint allows; and this is a feature of languages that ought to be explored.

[3]Strictly speaking, according to this definition w can never be non-conservative in L if the w-free part of L is already inconsistent; so perhaps we should rephrase the definition in terms of what is licensed by consistent fragments of L.

[4]See his [**Bra94**]. It is not essential to the general point I am making that his account of pejorative terms be correct.

Bibliography

[AB75] Alan Ross Anderson and Nuel D. Belnap. *Entailment*, volume 1. Princeton University Press, Princeton, 1975.

[Acz88] Peter Aczel. *Non-Well-Founded Sets*. CSLI Lecture Notes, Stanford, 1988.

[Aus50] J. L. Austin. Truth. In *Proceedings of the Aristotelian Society*, volume 24, 1950.

[BE87] Jon Barwise and John Etchemendy. *The Liar*. Oxford University Press, 1987.

[Bel62] Nuel D. Belnap. Tonk, plonk and plink. *Analysis*, 22:130–4, 1962.

[Bla86] Steven Blamey. Partial logic. In Dov Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic*, volume 3. D. Reidel, Dordrecht, 1986.

[Bra94] Robert Brandom. *Making It Explicit*. Harvard University Press, Cambridge, Mass., 1994.

[Bur79] Tyler Burge. Semantical paradox. *Journal of Philosophy*, 10:169–198, 1979. Page references are to the reprinting in [**Mar84**].

[Bur82] Tyler Burge. The liar paradox: Tangles and chains. *Philosophical Studies*, 41:353–66, 1982.

[Bur86] John P. Burgess. The truth is never simple. *Journal of Symbolic Logic*, 51:663–81, 1986.

[Chi79] Charles Chihara. The semantic paradoxes: A diagnostic investigation. *Philosophical Review*, 88:590–618, 1979.

[Dav67] Donald Davidson. Truth and meaning. *Synthese*, 17:304–323, 1967.

[Fef82] Solomon Feferman. Toward useful type-free theories I. *Journal of Symbolic Logic*, 49:75–111, 1982.

[Gai88] Haim Gaifman. Operational pointer semantics: Solution to self-referential puzzles I. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 43–59. Morgan Kauffman, Los Angeles, 1988.

[Gai92] Haim Gaifman. Pointers to truth. *Journal of Philosophy*, 89:223–261, 1992.

[GB93] Anil Gupta and Nuel D. Belnap. *The Revision Theory of Truth*. MIT Press, Cambridge, Mass., 1993.

[GCB75] Dorothy L. Grover, Joseph Camp, and Nuel D. Belnap. A prosentential theory of truth. *Philosophical Studies*, 27:73–125, 1975.

[Gri75] H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics III: Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[Gup82] Anil Gupta. Truth and paradox. *Journal of Philosophical Logic*, 11:1–60, 1982.

[Her66] Hans Herzberger. The logical consistency of langauge. In J. A. Emig, J. T. Fleming, and H. M. Popp, editors, *Language and Learning*. Harcourt, Brace and World, New York, 1966.

[Her67] Hans Herzberger. The truth-conditional consistency of natural languages. *Journal of Philosophy*, 64:29–35, 1967.

[Hor90] Paul Horwich. *Truth*. Basil Blackwell, Oxford, 1990.

[Kri75] Saul A. Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72:690–716, 1975. Page references are to the reprinting in [**Mar84**].

[Kri82] Saul A. Kripke. *Wittgenstein on Rules and Private Language*. Harvard University Press, Cambridge, Mass., 1982.

[Mar84] Robert L. Martin. *Recent Essays on Truth and the Liar Paradox*. Oxford University Press, 1984.

[McG91] Vann McGee. *Truth, Vagueness, and Paradox*. Hackett, Indianapolis, 1991.

[Mos74] Yiannis Moschovakis. *Elementary Induction on Abstract Structures*. North-Holland, Amsterdam, 1974.

[MW75] Robert L. Martin and P. W. Woodruff. On representing 'true-in-L' in L. *Philosophia*, 5:217–221, 1975.

[Par74] Charles Parsons. The liar paradox. *Journal of Philosophical Logic*, 3:381–412, 1974. Page references are to the reprinting in [**Mar84**].

[Par90] Terrence Parsons. True contradictions. *Canadian Journal of Philosophy*, 20:335–354, 1990.

[Pri60] A. N. Prior. The runabout inference ticket. *Analysis*, 21:38–9, 1960.

[Pri79] Graham Priest. The logic of paradox. *Journal of Philosophical Logic*, 8:219–41, 1979.

[Pri84] Graham Priest. The logic of paradox revisited. *Journal of Philosophical Logic*, 12:153–79, 1984.

[Rus08] Bertrand Russell. Mathematical logic as based on the theory of types. *Americal Journal of Mathematics*, 30:222–262, 1908.

[Soa97] Scott Soames. *Understanding Truth*. Oxford University Press, 1997.

[Str50] P. F. Strawson. Truth. In *Proceedings of the Aristotelian Society*, volume 24, pages 129–156, 1950.

[Tak75] Gaisi Takeuti. *Proof Theory*. North-Holland, Amsterdam, 1975.

[Tap92] Jamie Tappenden. *Vagueness and Truth*. PhD thesis, Princeton University, 1992.

[Tar35] Alfred Tarski. Der wahrheitsbegriff in den formalisierten sprachen. *Studia Logica*, 1:261–405, 1935. English translation by J. H. Woodger, The concept of truth in formalized languages, in [**Woo56**].

[Tar44] Alfred Tarski. The semantic conception of truth. *Philosophy and Phenomenological Research*, 4:341–375, 1944.

[Vis89] Albert Visser. Semantics and the liar paradox. In Dov Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic*, volume 4, pages 617–706. D. Reidel, Dordrecht, 1989.

[Woo56] J. H. Woodger, editor. *Logic, Semantics, Metamathematics*. Oxford University Press, 1956.