TRUTH AND ALETHEIC PARADOX

by

Kevin Scharp

AB, Washington University, 1995

MA, University of Wisconsin-Milwaukee, 1998

Submitted to the Graduate Faculty of

Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2005

UNIVERSITY OF PITTSBURGH

FACULTY OF ARTS AND SCIENCES

This dissertation was presented

by

Kevin Scharp

It was defended on

9 May 2005

and approved by

Anil Gupta, Distinguished Professor of Philosophy, University of Pittsburgh

John McDowell, University Professor of Philosophy, University of Pittsburgh

Hartry Field, Professor of Philosophy, New York University

Dissertation Director: Robert Brandom, Distinguished Service Professor, University of Pittsburgh

TRUTH AND ALETHEIC PARADOX

Kevin Scharp, PhD

University of Pittsburgh, 2005

My objective is to provide a theory of truth that is both independently motivated and compatible with the requirement that semantic theories for truth should not demand a substantive distinction between the languages in which they are formulated and those to which they apply. I argue that if a semantic theory for truth does not satisfy this requirement, then it is unacceptable. The central claim of the theory I develop is that truth is an inconsistent concept: the rules for the proper use of truth are incompatible in the sense that they dictate that truth both applies and fails to apply to certain sentences (e.g., those that give rise to the liar and related paradoxes). The most significant challenge for a proponent of an inconsistency theory of truth is producing a plausible theory of inconsistent concepts. On the account I provide, inconsistent concepts are confused concepts. A concept is confused if, in employing it, one is committed to applying it to two or more distinct types of entities without properly distinguishing between them; that is, an employer of a confused concept thinks that two or more distinct entities are identical. I propose a semantic theory for predicates that express confused concepts, and a new many-valued relevance logic on which the semantic theory depends. This semantic theory serves as the basis for my theory of inconsistent concepts. Given this account of inconsistent concepts and my claim that truth is inconsistent, I am committed to the view that truth is confused. I use the semantic theory for confused predicates as a semantic theory for truth. On the account I advance, a proper theory of truth requires a distinction between several different types of truth predicates. I propose an account of each truth predicate, and I advocate using them as consistent replacements for the concept of truth. The result is a team of concepts that does the work of the inconsistent concept of truth without giving rise to paradoxes.

TABLE OF CONTENTS

LIST OF FIGURES

PREFACE

> "Smokey, this is not 'Nam. This is bowling; there are rules."
> Walter Sobchak, *The Big Lebowski*

Numerous people have helped me with the ideas presented in this dissertation—I hereby thank all of them. Several people have had a significant influence on its style and content. In particular, I would like to thank Seth Ard, Brad Cokelet, Cian Dorr, Arthur Fine, Peter Guildenhuys, Graham Hubbs, John Koethe, Cristina Lafont, Steve Leeds, John Morrison, Jamsheed Siyar, Neil Tennant, and Steve Yablo for stimulating conversations on truth and related topics. In addition, Cian Dorr, Graham Hubbs, Peter Guildenhuys, and John Morrison read entire drafts of the dissertation—their comments, objections, and advice have been much appreciated.

I am especially grateful to the members of my dissertation committee. John McDowell's penetrating comments and enlightening conversations have been an inspiration to me over the past few years—he is truly a philosopher's philosopher. Despite the fact that Joseph Camp was not directly involved in the final version, his influence is obvious on almost every page of this document—I hope to be able to discuss it with him someday. A barrage of constructive criticism from Anil Gupta led me to improve many aspects of my views, my arguments, and my presentation. Hartry Field offered plenty of insightful comments; it has been a pleasure to work through these issues with him. My biggest intellectual debt is, without question, to Robert Brandom. He has been an ideal advisor. His ideas motivated me long before I came to Pittsburgh, and I will value his encouragement and his advice long after I leave. I could not ask for a better example of a thinker, a scholar, and a teacher.

I also want to express my appreciation to the members of my family for their love and for their role in helping me achieve this goal. Finally, I want to thank Alison Kerr for her presence, her passion, her insight, and her unwavering support; my life would be unrecognizable without her.

INTRODUCTION

The contemporary literature on truth in the analytic tradition divides into two groups. The first contains writings on the nature of truth, while the second contains writings on the logic of truth. Those in the former category address such questions as ‘what is truth?’, discussing, e.g., which entities can be true or false, whether truth is a substantive concept or a logical one, and how best to account for the ways we use truth predicates. Those in the latter category aim to describe the logical principles governing languages that contain truth predicates, focusing on the liar paradox and related phenomena.

While each of these traditions is well developed and thriving, there is little interaction between them. I argue that this lack of communication is unacceptable. A proper account of the nature of truth is the key to understanding the liar paradox, while the lessons learned from investigating the logic of truth are crucial to understanding what truth is. My dissertation contains a theory of truth that draws on the insights of both traditions.

I begin by arguing for a condition on any theory of truth. The condition is intended to capture the intuition that a theory of truth that offers an approach to the liar paradox should not require a substantive distinction between the language in which it is formulated and the languages to which it applies. I provide an argument for the claim that semantic theories for truth that require such a distinction are unacceptable. The argument turns on the claim that these theories cannot be applied successfully to natural languages. This requirement on theories of truth is difficult to meet—of the dozens of theories proposed in the last forty years, only a few even purport to meet it. Thus, the requirement serves as an effective criticism of a wide range of theories from each tradition.

The objective of my dissertation is to provide a theory of truth that is both independently motivated and compatible with the requirement that semantic theories for truth should not demand a substantive distinction between the languages in which they are formulated and those to which they apply. The theory of truth I develop satisfies both these demands: it is capable of accounting for our practice of using truth predicates, and it is expressible in the languages to which it applies. The central claim of this account is that truth is an inconsistent concept: the rules for the proper use of truth are incompatible in the sense that they dictate that truth both applies and fails to apply to certain sentences (e.g., those that give rise to the liar and related paradoxes). The most significant challenge for a proponent of an inconsistency theory of truth is producing a plausible theory of inconsistent concepts. Accordingly, I first construct a theory of inconsistent concepts, and then I apply it to truth.

On the account I provide, inconsistent concepts are confused concepts. A concept is confused if, in employing it, one is committed to applying it to two or more distinct types of entities without properly distinguishing between them; that is, an employer of a confused concept thinks that two or more distinct entities are identical. (An example popularized by Hartry Field is the Newtonian concept of mass.) I propose a semantic theory for predicates that express confused concepts, and a new many-valued relevance logic on which the semantic theory depends. This semantic theory serves as the basis for my theory of inconsistent concepts.

Given my account of inconsistent concepts and my claim that truth is inconsistent, I am committed to the view that truth is confused. Accordingly, I use the semantic theory for confused predicates as a semantic theory for truth. On the account I advance, a proper theory of truth requires a distinction between several different types of truth predicates. I propose an account of each truth predicate, and I advocate using them as consistent replacements for the concept of truth. An important feature of this account is that it permits a generic truth predicate, which allows our practice of using ‘true’ to go on without much interruption even though our inconsistent concept of truth has been replaced with a group of consistent ones. The generic truth predicate can be used in most conversational situations without having to consider which specific concept of truth is appropriate; only in very specialized situations does one need to distinguish between the various types of truth. The result is a team of concepts that does the work of the inconsistent concept of truth without giving rise to paradoxes.

I set out to provide both a compelling argument against theories of truth that require a substantive distinction between the language in which they are formulated and those to which they apply, and an alternative theory of truth that is independently motivated. Chapters One and Two, address the first task. Because I advocate a theory of truth on which truth is an inconsistent concept, the second task requires a theory of inconsistent concepts. In Chapter Three, I motivate the claim that truth is an inconsistent concept. Chapters Four, Five, and Six contain the theory of inconsistent concepts, and in chapter Seven, I apply this theory to truth.

Chapter One

Most theories of truth that address the liar paradox must be formulated in a language that is expressively richer than the languages to which they apply. Although some philosophers have objected to theories with this feature, no one has formulated a detailed criticism of them. I set up such a criticism in Chapter One by developing a system of concepts and distinctions capable of characterizing theories with this feature in a sufficiently precise way. Of these concepts, the most important is internalizability. A semantic theory for truth is *internalizable* for a language if and only if there exists an extension of the language such that (i) the theory is expressible in that extension, and (ii) the theory assigns meanings to all the sentences of that extension that express the concept of truth.

Chapter Two

This chapter is devoted to arguments for two internalizability requirements. First, a semantic theory should be internalizable for every language; a semantic theory for some concept that fails to meet this requirement fails to adequately explain that concept. Second, a semantic theory for truth that applies to a natural language should be internalizable for that natural language. Unless a semantic theory for truth meets this requirement, it will not be able to describe the natural language in question. Of course, any theory that satisfies the first requirement satisfies the second. However, providing a separate defense enables one to handle the objection that even if a semantic theory for truth is not internalizable for every language, it will still work well for natural languages. Very few, if any, semantic theories for truth are internalizable for natural languages; hence, the internalizability requirements serve as powerful criticisms of most theories of truth.

Chapter Three

In this chapter, I explain why it is so difficult to construct a semantic theory for truth that is internalizable for natural languages. I show that theories of truth that treat truth as a consistent concept must inevitably resort to a substantive distinction between object language and metalanguage to maintain consistency. Such theories do not satisfy the internalizability requirement I defend in Chapters One and Two. In the remainder of the chapter, I provide further defense of the view that truth is an inconsistent concept.

Chapter Four

The most significant challenge for an inconsistency theory of truth is a plausible theory of inconsistent concepts. In Chapter Four, I present the outlines of such a theory. I argue that the constitutive rules for an inconsistent concept are incompatible, where a concept's constitutive rules are those that a person who employs that concept is committed to following by virtue of his or her employment of that concept. That is, when one employs a concept, one is obligated to follow its constitutive rules. In the case of inconsistent concepts, the constitutive rules stipulate that the concept should both apply and fail to apply to some object or objects. Such concepts are perplexing. It is unclear how we should understand them, whether sentences that express them are true or false, and how we should evaluate arguments that contain such sentences. My views on these issues all stem from my commitment to explain inconsistent concepts in terms of confusion. A person is confused if she thinks that two or more distinct entities are identical. One can be confused about a variety of things, but I focus on cases in which someone is *conceptually* confused. Newtonian mass serves as good example of conceptual confusion. Today we know that there are two "kinds" of mass (proper mass and relativistic mass), but before the advent of relativity, people who employed the concept of Newtonian mass assumed that there was a unique property, mass, that objects possess. Hence, the concept of Newtonian mass is a confused concept.

Chapter Five

Given that I explain inconsistent concepts in terms of confused concepts, I require a theory of confused concepts. In this chapter, I describe and defend one such theory that takes Joseph Camp's theory of confused names as its starting point. According to this theory, confusion is not to be understood as a mental state, but rather as a certain status. Camp's theory of confusion employs a four-valued logic to evaluate arguments that contain confused expressions. The logic has four "epistemically interpreted" semantic values: told true, told false, told neither, and told both. One can show that the inferences validated by this logic are those that are validated by the logic of first-degree entailments (a relevance logic). To employ the four-valued logic, one must first determine the components of the confused concept, which are the distinct entities thought to be identical by the confused person (e.g., the components of Newtonian mass are proper mass and relativistic mass). One then generates semantic values for sentences that contain confused expressions by substituting the component expressions into a confused sentence and evaluating the result for truth (e.g., one determines whether the sentence that results from substituting 'proper mass' for 'mass' in the confused sentence in question is true; one does the same for 'relativistic mass'). One can then use the four-valued logic to evaluate arguments with confused expressions.

Chapter Six

In order for the theory of confusion described in Chapter Five to serve my needs, it must be extended in four ways; these extensions are the topic of Chapter Six. First, I advocate adding a certain conditional that renders the resulting logic equivalent to the sentential logic of relevant implication (R), and I advocate the standard first-order extension of R. Second, I present a family of many-valued logics that are intended to generalize the four-valued logic described in Chapter Five. The four-valued logic is what I call a *2-component logic* on the grounds that it allows for at most two epistemic perspectives (told true and told false). I generalize this to the n-component case so that the theory of confusion that employs these logics can handle confused concepts with more than two components. For example, if there had turned out to be three “kinds” of mass, then the concept of Newtonian mass would have had three components, and one would have needed to use a 3-component logic (which has six semantic values) to evaluate arguments whose sentences express this concept. Third, I construct a family of *partial* n-component logics, which include “told gappy” values. The result is that the theory of confusion can handle confused concepts whose components are partially defined, even if the components have different ranges of application. Finally, I combine the extended theory of confusion with a conceptual role semantics to arrive at a semantic theory for confused expressions. The chapter closes with a more detailed treatment of inconsistent concepts and a sequence of replies to objections.

Chapter Seven

In the final chapter, I apply the theory of inconsistent concepts to truth and arrive at an inconsistency theory of truth. Given the theory of inconsistent concepts I develop in Chapters Four, Five, and Six, I endorse the claim that truth is a confused concept. On the theory I propose, truth has six components: ascending weak truth, descending weak truth, ascending strong truth, descending strong truth, ascending dual truth, and descending dual truth. Ascending weak truth obeys the ascending truth rule (i.e., from $\vdash$ p infer $\vdash$ True(⟨p⟩)), but it obeys the descending truth rule (i.e., from $\vdash$True(⟨p⟩) infer$\vdash$ p) only for non-pathological sentences.

Descending weak truth obeys the descending weak truth rule, but it obeys the ascending weak truth rule only for non-pathological sentences. Ascending versions of strong truth and dual truth are defined in terms of ascending weak truth, while descending versions of strong truth and dual truth are defined in terms of descending weak truth. Ascending (descending) strong truth has the same extension as ascending (descending) weak truth, but is completely defined on truth-apt sentences; ascending (descending) dual truth has the same anti-extension as ascending (descending) weak truth, but is completely defined on truth-apt sentences. I use these six components as the basis for the semantic theory for truth I develop; it uses a partial 6-component logic to evaluate arguments that express the inconsistent concept of truth.

In the end, I advocate replacing inconsistent concepts with consistent ones. For truth, the six component concepts constitute an ideal group of replacements. I demonstrate the importance of introducing a generic truth predicate whose extension is the union of the extensions of the six component concepts and whose anti-extension is the intersection of the anti-extensions of the six component concepts. The generic truth predicate allows our practice of using ‘true’ to go on without much interruption even though our inconsistent concept of truth has been replaced with a group of consistent ones. The generic truth predicate can be used in most conversational circumstances without having to consider which component concept of truth is appropriate. One needs to distinguish between the various types of truth only in very specialized situations (e.g., when providing a semantics for an expressively rich language).

Appendix A: Fragmentary Theories of Truth

Some theories of truth explain natural language truth predicates in terms of a group of restricted truth predicates; the extension of each restricted truth predicate is a proper subset of the extension of ‘true’. I call these *fragmentary theories of truth*. Examples of fragmentary theories of truth are Tarski’s theory, the disquotational version of deflationism, and most approaches to the liar paradox, including fixed-point theories, revision theories, and contextual theories. I argue that many fragmentary theories of truth are inconsistent with our intuitions about which assertions of truth ascriptions are warranted. Because these theories purport to describe the way language users actually use truth predicates, their failure to respect our intuitions on warranted assertibility renders them unacceptable.

Appendix B: Risky Business

It is relatively easy to construct versions of the liar paradox that employ empirical predicates. One surprising consequence of this claim is that whether a sentence token counts as paradoxical can depend on empirical facts that are independent of its syntactic and semantic features. In other words, paradoxicality does not supervene on the syntactic and semantic features of sentence tokens. This fact has profound consequences for the study of truth. In particular, I argue that it casts doubt on the claim that utterances of paradoxical sentences are not assertions, the claim that propositions are primary truth bearers, minimalist theories of truth aptness, a prominent version of deflationism, and contextual approaches to the liar paradox.

Appendix C: Revision and Revenge

The revision theory of truth is currently one of the three most prominent approaches to the liar paradox (the others being fixed-point theories and contextual theories). Although the revision theory must be formulated in a language that is expressively richer than those to which it applies, its proponents claim that it does not fall prey to revenge paradoxes. I argue that it does face a revenge paradox, and this revenge paradox casts doubt on both its claim to be an acceptable approach to the liar paradox and its prospects for applying to natural languages.

Appendix D: Purportedly Internalizable Semantic Theories for Truth

Several semantic theories for truth have appeared in the last few decades, and the supporters of these theories claim that they do not require a substantive distinction between object language and metalanguage. I compare and contrast these theories, and I evaluate the extent to which they are internalizable for a single language and for natural languages.

Appendix E: Theories of Inconsistent Concepts

I compare the theory of inconsistent concepts developed in Chapters Four, Five, and Six with several other theories designed to handle inconsistent concepts.

1.0 Internalizable Semantic Theories

1.1 Introduction

In this chapter and the next one, my aim is to provide a detailed argument for a claim that appears frequently in philosophical discussions of truth, but has, in my view, never been adequately defended. The claim to which I allude is: a semantic theory for truth that must be formulated in a language that is expressively richer than the languages to which it applies is unacceptable. This issue arises in the literature on truth because the vast majority of semantic theories for truth do not apply to the languages in which they are formulated. Thus, if the arguments I provide below are successful, then they constitute a refutation of the vast majority of semantic theories for truth.

In Chapter One, I introduce some terminology in an attempt to capture this intuition. In particular, for a given language, I distinguish between semantic theories that are internalizable for that language and those that are essentially external for that language. I claim that semantic theories that are internalizable for every language can be expressed in the languages to which they apply, while semantic theories that are essentially external for every language are those that require formulation in languages that are essentially richer than the languages to which they apply. In Chapter Two, I argue for two internalizability requirements on semantic theories, the most important of which is that if a semantic theory for truth can be successfully applied to a natural language, then that semantic theory is internalizable for that language.

Before presenting my account of internalizable semantic theories, I want to provide the reader with some background on truth and the liar paradox in an effort to motivate the distinctions I make in this chapter.[1] Let me begin by stipulating that a *theory of truth* is any theory that specifies some aspect of the nature of truth, and a *semantic theory for truth* is a theory that provides the meanings of the sentences in a particular language that express the concept of truth. I discuss these terms at length in section two, but these brief definitions should be adequate for now.

Any semantic theory for truth that applies to a language with minimal expressive resources must incorporate an approach to the liar paradox; otherwise, if it is remotely plausible, then it will almost certainly be inconsistent. The liar paradox involves sentences like the following (which I call a *liar sentence*):

(1): (1) is false.

There is an intuitively plausible argument whose conclusion is that (1) is both true and not true. The argument is depends on inference rules of classical logic and the *truth rules*: (i) the *ascending truth rule* (i.e., ⟨⟨p⟩ is true⟩[2] follows from ⟨p⟩), (ii) the *descending truth rule* (i.e., ⟨p⟩ follows from ⟨⟨p⟩ is true⟩), and (iii) the *substitution rule* (i.e., two names that refer to ⟨p⟩ are intersubstitutable in ⟨⟨p⟩ is true⟩ without changing its truth-value). On the one hand, if (1) is true, then ‘(1) is false’ is true (by substitution). If ‘(1) is false’ is true, then (1) is false (by descending). Thus, if (1) is true, then (1) is false. On the other hand, if (1) is false, then ‘(1) is

[1] In these introductory remarks, I have sacrificed rigor for accessibility. I hope that readers who are familiar with the formal presentations of these issues will be patient with my attempts to introduce them to those readers who are not.

[2] ‘⟨’ and ‘⟩’ are angle quotes; ‘p’ serves as a sentential variable that can be replaced by a sentence, and ‘⟨p⟩’ is the quote-name of such a sentence. See McGee (1991, 2000) for this usage (McGee uses different symbols). I also use ‘p’ as a logical constant (e.g.: p is true). Note that these uses are distinct: an occurrence of ‘p’ cannot be both a sentential variable and a constant. Corner quotes, ‘⌜’ and ‘⌝’, are used in conjunction with constants. For example, if ‘p’ and ‘q’ are names of sentences, then ‘⌜p ∧ q⌝’ is the name of the sentence that results from placing sentence p and sentence q on opposite sides of ‘∧’.

false' is true (by ascending). If '(1) is false' is true, then (1) is true (by substitution). Thus, if (1) is false, then (1) is true. Therefore, (1) is true if and only if (1) is false. It follows that (1) is both true and false. The paradox is that from intuitively plausible assumptions via intuitively plausible inferences, one can derive an intuitively unacceptable conclusion. Anyone who endorses a semantic theory for truth that applies to a language with sentences like (1) must reject one of the premises, reject one of the inferences, or accept the conclusion.

As I mentioned, most semantic theories for truth do not apply to the languages in which they are formulated. The reason turns out to be linked to the liar paradox. One of the most frustrating and ubiquitous features of approaches to the liar paradox is the presence of revenge paradoxes. A *revenge paradox* for a given semantic theory for truth involves a sentence (a *revenge liar sentence*) that is similar to the liar sentence, but the sentence in question expresses a key concept used by the semantic theory to classify the liar sentence. One can derive a contradiction using the revenge liar from assumptions and inferences that are compatible with the semantic theory in question. For example, a popular approach to the liar paradox stipulates that liar sentences are truth-value gaps (i.e., they are neither in the extension nor in the anti-extension of truth). A semantic theory for truth that incorporates this approach classifies sentences of a language as true, false, or gappy (given certain information about the sentences and the world in which they are evaluated). The revenge paradox for such a semantic theory involves the following sentence:

(2): (2) is either false or a truth-value gap.

There is an intuitively plausible argument whose conclusion is that (2) is both true and not true.[3] The liar paradox takes its revenge on the semantic theory in question by giving rise to a new

[3] If (2) is either false or a truth-value gap, then '(2) is either false or a truth-value gap' is true. If '(2) is either false or a truth-value gap' is true, then (2) is true. Hence, if (2) either false or a truth-value gap, then (2) is true.

paradox that uses a key concept of the semantic theory (in this example it is the concept of a truth-value gap). Victims of revenge paradoxes commonly restrict their semantic theories for truth so that the theories do not apply to languages with the resources to construct revenge paradoxes. Because the revenge paradox incorporates key terms of the semantic theory itself, this restriction prevents the semantic theory from applying to the language in which it is formulated.[4]

A related phenomenon forces the same sort of restriction. This phenomenon is sometimes confused with the revenge paradox phenomenon, and it does not seem to have a commonly used name. I call it the *self-refutation problem*. The problem is that most semantic theories for truth imply that paradoxical sentences (e.g., (1) and (2)) are not true.[5] However, one can construct a paradoxical sentence that attributes non-truth to itself. Thus, a semantic theory for truth that implies that paradoxical sentences are not true will have a paradoxical (and hence, an untrue) sentence as a consequence. Hence, this semantic theory for truth implies that one of its consequences is untrue. Therefore, this semantic theory is self-refuting. The self-refutation problem occurs when a semantic theory implies that a certain sentence is not true and the sentence in question is a token of the same type as the consequence of the theory. In other words, some paradoxical sentences seem to mimic what a semantic theory says about them. For example, consider again the semantic theory for truth that incorporates the truth-value gap

Likewise, if (2) is true, then '(2) is either false or a truth-value gap' is true. If '(2) is either false or a truth-value gap' is true, then (2) is either false or a truth-value gap. Hence, if (2) is true, then (2) is either false or a truth-value gap. Consequently, (2) is true if and only if (2) is either false or a truth-value gap. Therefore, (2) is both true and either false or a truth-value gap (a contradiction—if we assume that a sentence that is either false or a truth-value gap is not true).

[4] See van Fraassen (1968), Parsons (1974), Kripke (1975), Burge (1979a), Priest (1987), McGee (1991), Simmons (1993), Gupta and Belnap (1993), Gupta (2000), McDonald (2000), and Field (2003a, 2003b) for discussion.

[5] That is, most semantic theories for truth imply that paradoxical sentences are untrue. The 'not' in the sentence on which this footnote is a comment should be read as exclusion negation.

approach to the paradoxical sentences. The self-refutation problem for this semantic theory concerns the following sentence:

(3): (3) is not true.

There is an intuitively plausible argument whose conclusion is that (3) is both true and not true. Thus, the semantic theory in question should classify (3) as a truth-value gap. However, if (3) is a truth-value gap, then (3) is not true. Consequently, the semantic theory in question implies that (3) is not true. However, that is just the content of (3) itself. Therefore, the semantic theory in question has '(3) is not true' as a consequence, and it implies that '(3) is not true' is not true. Thus, it is self-refuting—it implies that one of its consequences is untrue. The standard response to the self-refutation problem is to restrict one's semantic theory to languages in which such sentences cannot be formulated. Semantic theories with this restriction are not self-refuting because they do not have any consequences for sentences like (3). Thus, they do not imply that their own consequences are untrue. Like a revenge paradox, a self-refutation problem for a semantic theory pertains to a sentence that expresses a concept employed by the theory itself; thus, restricting the theory to avoid the self-refutation problem prevents it from applying to the language in which it is formulated.[6]

Several philosophers have claimed that restricting a semantic theory for truth so that it does not give rise to revenge paradoxes or self-refutation problems is unacceptable. For example, Vann McGee proposes the "integrity of language" requirement, which states, "[i]t must be possible to give the semantics of our language within the language itself," (McGee 1991: 159). McGee comments on his requirement that it "is intended to hold open the possibility that the methods we develop can be applied to natural languages. If in developing the theory of truth

[6] See Burge (1979), Priest (1987), Simmons (1993), Gupta (2000), and Glanzberg (2003) for more on the self-refutation problem.

for a language, we required the services of an essentially richer metalanguage, that possibility would be closed off. … [It] makes it reasonable to hope that our methods can be used to get a semantics of a natural language," (ibid.) William Reinhardt expresses a similar sentiment in the following passage:

> Let us suppose, as I believe is intuitively correct, that one of the primary features of [truth] is that it is one notion: in particular it does not split into some hierarchy of notions. … Let us explain that the truth predicate of our formal language (call the language L) is intended to be taken in the sense of our preexisting informal notion of truth. … Unless we are prepared to entertain splitting the notion of truth, we are forced to admit that the metalanguage is included in the object language. If the formal language is to provide an adequate explication of the informal language that we use, it must contain its own metalanguage. I take it that this is in fact a desideratum for success in formulating a theory of truth, (Reinhardt 1986: 227-228).

Both McGee and Reinhardt suggest that if a semantic theory for a language requires a richer language for its formulation, then the theory cannot be applied successfully to natural languages.[7] Two of my goals in the first two chapters are to formulate this intuition in a sufficiently precise way and to argue for it.

In section one, I consider several arguments for the intuition voiced by McGee and Reinhardt and reject each of them. Section two contains a definition of internalizability, which is the central concept I use to capture this intuition, and a discussion of several related concepts. To illustrate these concepts, I apply them to Kripke's semantic theory for truth. I then turn in Chapter Two to formulating several requirements on semantic theories in general and semantic theories for truth in particular that are similar to the above intuition. The rest of Chapter Two

[7] Kearns, Priest, Simmons, and Martin express similar sentiments as well; see Kearns (1970), Priest (1987), Simmons (1993), and Martin (1997). See also Field's discussion in Field (2003a, 2003b). In the following passage, Brandom imposes a similar requirement on his theory of meaning: "One of the criteria of adequacy that has guided the project from the outset is that it be possible to elaborate the model of discursive practice to the point where it is characterized by just this sort of expressive completeness. This means that the model reconstructs the expressive resources needed to describe the model itself," (Brandom 1994: 641). However, there is a subtle difference between this requirement and the one imposed by McGee and Reinhardt; see below on the distinction between theories of meaning and meaning-theories and footnote 35 for more on the difference.

contains arguments for two of these requirements. The first argument pertains to any semantic theory whatsoever, while the second concerns semantic theories for truth that apply to natural languages. Chapter Two closes with a discussion of Tarski's indefinability theorem and related results.

1.2 Previous Proposals

In this section, I consider three arguments for the claim that semantic theories for truth that cannot be formulated in the languages to which they apply are unacceptable. The first is an argument of McGee's, which depends on a commitment to naturalism. The second is based on certain views about our capacity to comprehend natural languages. The third depends on claims about the expressive capacity of natural languages; it is explicitly formulated by Anil Gupta (in an effort to discredit it) as part of his defense of the revision theory of truth (which does not apply to languages in which it is formulated).

1.2.1 The Naturalism Argument

Although philosophers have used 'naturalism' as the name of several different doctrines, McGee's usage is common.[8] McGee's naturalist believes that the methods and concepts of science are amenable to humans and our activities, which include natural languages. Consequently, the naturalist believes that it is possible to provide a theory of human language that encompasses all human languages, even the one that is used to formulate the theory. Hence,

[8] See King (1994) for discussion of the types of naturalism.

naturalists ought to reject semantic theories that do not apply to the languages in which they are formulated.[9]

McGee assumes that there is something all languages have in common and that, whatever this is, it calls out for a scientific explanation. One response to the naturalism argument is to deny that the lack of a unified scientific theory of language poses a threat to naturalism. Perhaps languages can be explained only one by one, or maybe it is only fragments of natural languages that can be explained as coherent units. One could endorse such a piecemeal explanation and maintain naturalism as long as the explanation employed only scientifically acceptable concepts. There is evidence that McGee himself should accept piecemeal explanations of language as naturalistically acceptable because he uses this claim as an objection to Field's criticism of Tarski. Field argues that Tarski's truth definition is not acceptable to naturalists because it employs an enumerative account of reference and predication.[10] McGee defends Tarski by claiming that enumerative accounts are naturalistically acceptable, and he cites lepidoptery as an example of a science that is enumerative. He even argues that a piecemeal account of language should be expected because of the conventional character of languages.[11] Hence, McGee provides a good reason to think that a piecemeal explanation of language could be legitimate in the eyes of a naturalist. If this view is correct, then there is no reason to think that an appeal to naturalism justifies the claim that semantic theories should apply to the languages in which they are formulated.

[9] McGee (1991: ix; 1994: 628-9).
[10] Field (1972) and Tarski (1933).
[11] McGee (1991: 83-86).

1.2.2 The Comprehensibility Argument

A different argument for the intuition is based on the comprehensibility of natural languages. Many humans have the ability to speak, write, hear, and read natural languages. Let anyone who has one of these capacities for a particular natural language be said to *comprehend* that natural language. According to some views of comprehension, human comprehension is theory-based in the sense that someone who comprehends a particular natural language uses a theory that prescribes the use of the elements of that language.[12] Someone who comprehends, say, English, could simply state the theory on which his comprehension is based and he would have a semantic theory for English. Hence, English (and any other natural language that can be comprehended) must be able to express its own semantic theory.[13]

There are several weak points in the argument. One objection is that it relies on the assumption that human linguistic competence is based on the acceptance or knowledge of a semantic theory.[14] It seems to me that this assumption is linked to a certain view on the nature of linguistic competence. Fodor expresses it adequately in this passage: “To describe a language is to formulate the rules which are internalized by speakers when they learn the language and applied in speaking and understanding it,” (Fodor 1964: 198). Davidson, in particular, has criticized the claim that comprehending a language is internalizing a set of rules for uttering and

[12] Proponents of the *theory theory* and those who advocate the *simulation theory* have been engaged in a prominent debate for several decades now about how humans come to know about one another’s mental states. See the papers in Carruthers and Smith (1996) for an overview of the debate. A similar issue concerns the way in which humans come to know about the semantic features of words and sentences, but this issue receives considerably less attention than the debate about mental states. It seems to me that someone who endorses the comprehensibility argument would have to accept a variant of the theory theory for linguistic competence.

[13] Gupta formulates this argument for the purposes of attacking it. He says that it is the most common such argument and attributes it to Simmons, but I am not convinced that Simmons advocates such an argument. In the passage of Simmons (1993) that Gupta quotes, Simmons argues that natural languages are semantically universal because they are semantically closed, not that anyone with the capacity to comprehend a language can state the theory underlying that capacity. Regardless of who advocates it, it seems to be a relatively intuitive argument that many people with whom I have discussed these issues find compelling. See Gupta (1997: 440 n. 22) and Simmons (1993: 15).

[14] Gupta presents this objection in his discussion of the argument. See Gupta (1997: 440-1).

interpreting its words and sentences.[15] I am not interested in engaging with this debate here. I do want to point out that the comprehensibility argument seems to have this controversial assumption.

Another problem with it is that it might be that a human can comprehend a particular natural language and that this capacity is theory-based without it being the case that the human in question can express the theory he uses to comprehend the language *in that language*. Perhaps comprehension of English is based on a theory that can be expressed in the language of thought, but is not expressible in English.

1.2.3 Expressive Capacity Arguments

Some critics of semantic theories that do not apply to languages in which they are formulated argue that such semantic theories cannot successfully describe natural languages. There are several ways to make such an argument, but they all have the same form: (i) natural languages have certain expressive resources, (ii) a semantic theory that cannot apply to the languages in which it is formulated cannot apply to languages with those expressive resources, thus, (iii) a semantic theory that cannot apply to the languages in which it is formulated cannot apply to natural languages.

The expressive resources cited differ depending on who formulates the argument, but one popular choice is semantic self-sufficiency.[16] A language is *semantically self-sufficient* if and only if it can express its own semantic theory (i.e., a semantic theory that assigns a meaning to

[15] Davidson (1986).
[16] Others are universality and semantic closure (see section 1.4 for discussion).

every sentence of the language).[17] Gupta, in an effort to defend the revision theory of truth, formulates this version of the expressive capacity argument:

(a) It is possible to provide a semantic description of a natural language L.

(b) A semantic description of L must be expressible in L (i.e., L is semantically self-sufficient).

(c) The revision theory must be formulated in a language that is expressively richer than the one it describes.

(d) The revision theory is not suitable for L (from (a), (b) and (c)).

∴ (e) The revision theory fails to explain truth in L (from (d)).

Gupta then attacks this argument by claiming that we have no reason to believe that natural languages are semantically self-sufficient.[18]

It seems to me that the reasoning Gupta reconstructs is common among those who criticize semantic theories that require formulation in languages that are expressively richer than the ones to which they apply. Although I agree with Gupta that we have no reason to think that natural languages are semantically self-sufficient, I find another problem with this argument. It is improper to infer from the claim that L *can* express a semantic theory for L to the claim that, to be acceptable, a semantic theory for L *must* be expressible in L. It is possible that there is a perfectly good semantic theory for a semantically self-sufficient language that is not expressible in that language (of course, it would not be the semantic theory for that language that is expressible in it). Hence, the inference from (a), (b), and (c) to (d) is invalid. Therefore, even if one could show that natural languages are semantically self-sufficient, that would not, by itself,

[17] That is, it must be able to express a theory that *correctly* assigns a meaning to every sentence of the language. There must be some sort of correctness clause in the definition or else most any language would count as semantically self-sufficient by virtue of expressing the claim that all its sentences are meaningless or that all its sentences mean that Roquefort is yummy.

[18] Gupta (1997: 437).

constitute a criticism of semantic theories that cannot apply to the languages in which they are formulated (the same point undermines the comprehensibility argument as well).

A different sort of expressive capacity argument can be formulated for an individual semantic theory. For example, natural languages (e.g., English) can express the concept of a truth-value gap. A particular semantic theory for truth that employs the concept of a truth-value gap cannot apply to languages that can express the concept of a truth-value gap. Thus, this semantic theory does not apply to natural languages. Although this type of argument can serve as a criticism of an individual semantic theory, it does not justify the claim that any semantic theory for truth should apply to the languages in which it is formulated. To get that result, one would have to claim that natural languages are universal in the sense that they can express any concept whatsoever. That leads one back to the first type of expressive capacity argument.

To sum up: although the intuition expressed by McGee and Reinhardt could be used as a powerful critique of most theories of truth, it lacks justification. In particular, none of the arguments surveyed in this section gives us good reason to accept it.[19]

1.3 Internalizable Semantic Theories

In this section, I present a conceptual framework designed to capture the intuition that semantic theories that do not apply to languages in which they are formulated are unacceptable. At the center of this framework is the concept of internalizability. In the first subsection, I define

[19] Another argument for the same conclusion is that a semantic theory that does not apply to the languages in which it is formulated is self-refuting; see Fitch (1946) and Simmons (1993: 58-61). It seems to me that once the restrictions on these theories are properly understood, it is clear that these theories are not self-refuting. See Chapter Three, section 3.2 for discussion.

‘internalizability’ and several related terms. In the second, I use Kripke’s semantic theory for truth to illustrate them.

1.3.1 Definitions

In order to expose and avoid certain difficulties, I first present and reject three suggestions for capturing the intuition voiced by McGee and Reinhardt. One suggestion is that semantic theories should be expressible in the languages to which they apply. However, this requirement is too strong. The above intuition is that semantic theories that *cannot* be expressed in the languages to which they apply are unacceptable. I do not want to say that a semantic theory that *happens to be* inexpressible in a language to which it applies is unacceptable. For example, assume that T is a semantic theory for counterfactual expressions that employs the concept of a possible world, but T need not be formulated in a language that is expressively richer than those to which it applies. Assume also that T applies to a language L that contains counterfactuals but does not have the vocabulary to express the concept of a possible world. In this case, T is not expressible in L. However, it is plausible to assume that L could be extended to a new language L′ such that T is expressible in L′ and T applies to L′. The first suggestion would reject theories like T even though they can be formulated in the very languages to which they apply.

A second suggestion is that a semantic theory should be expressible in an extension of the languages to which it applies. This condition is too weak because for any semantic theory and any language to which it applies, there exists an extension of that language in which the theory can be expressed, so long as the theory does not apply to the extended language. The intuition I am trying to capture is that a semantic theory should apply to the very languages in which it can be expressed.

As a final suggestion one might say that for a given semantic theory and a language to which it applies, the theory should both be expressible in an extension of that language and apply to that entire extended language. This characterization is better, but I want the account to work for semantic theories that focus on a single concept (e.g., a semantic theory for truth) as well as semantic theories for entire languages (e.g., a semantic theory for English).

The best formulation of the intuition is that a semantic theory that applies to a language is acceptable if and only if it is expressible in an extension of that language and applies to everything in that extension to which it is supposed to apply. A semantic theory of this type is internalizable for that language. The following is a more elaborate definition of internalizability:

> A semantic theory T that purports to specify the meanings of sentences that express a concept X is *internalizable for a language L* if and only if there exists an extension of L such that all the sentences that compose T can be translated into sentences that belong to the extension of L and T specifies the meanings of all the sentences of the extension of L that express X.

That is rather long-winded and contains many expressions whose meanings are unclear. In the interest of clarity, I discuss four aspects of this definition: semantic theory, language, expression, and application. Semantic theories and languages are objects (in a loose sense that includes abstract entities) and expression and application are relations between a semantic theory and a language.

I begin with 'semantic theory'. First, a *theory* is a set of declarative sentences all of which belong to a single language.[20] The term 'semantic theory' is tricky to define because it has been used in so many ways. I follow Dummett in distinguishing between meaning-theories

[20] This account of theoryhood is not without its problems. First, we usually think of theories as things that can be expressed in different languages. We find it natural to say that two physics textbooks, one written in English, the other written in French, both contain Newton's theory of mechanics. However, on my account, they contain two different theories. I attempt to defuse this problem by speaking of a theory and its translations into other languages. Furthermore, although philosophers (and many other people) use the term 'theory' quite often, it is rather difficult to say which sentences constitute a particular informal theory. I have no doubt that I would have trouble specifying the sentences that constitute Lewis's theory of natural laws or Davidson's paratactic theory of indirect discourse; see Lewis (1994) and Davidson (1968). Nevertheless, I stick with the idealization.

and the theory of meaning. For Dummett, *the theory of meaning* is the branch of philosophy that deals with the nature of meaning, while a *meaning-theory* is a particular theory that specifies the meanings of the words or sentences of a particular language or languages. I use the term 'a theory of meaning' in a Dummettian spirit to designate a theory that specifies the nature of meaning.[21] Theories of meaning provide necessary and sufficient conditions on meaning-theories. According to my usage, a semantic theory is a type of meaning-theory. In particular, a *semantic theory* is a theory that specifies the meanings of certain sentences that belong to some particular language or languages.[22] I use the locution 'semantic theory for X', where X is a placeholder for the name of a concept (e.g., a semantic theory for moral obligation, a semantic theory for truth). A semantic theory for X specifies the meanings of the sentences of certain languages that express the concept X (e.g., a semantic theory for truth specifies the meanings of sentences that express the concept of truth).[23] I also use the locution 'theory of X'; a *theory of X* is a theory that makes claims about the nature of X.[24]

A *language* is a function from sets of sentences (syntactic strings) to a set of sentential meanings. A *sublanguage* L_0 of a language L_1 is a language whose set of sentences is a subset of

[21] Dummett (1991: 20-22). See Peacocke (1981) for this use of 'a theory of meaning'.

[22] Dummett also uses the term 'semantic theory' but his account differs from mine. For Dummett, a semantic theory must specify the truth-value of each sentence in a given language (see Dummett 1991: 25, 33, 35). King (1994: 57) and Soames (2002: 97) define 'semantic theory' as I do.

[23] I briefly discuss the nature of concepts in section 4.2 of Chapter Four, but I prefer to accommodate a range of views on the nature of concepts.

[24] Although I do not make much of it, the distinction between a semantic theory for X and an X definition is important. An *X definition* provides the extension, the intension, or the sense of a word that expresses the concept X. (The *extension* of a predicate is the set of things of which the predicate is true; the *intension* of a predicate determines its extension across possible worlds, and the *sense* of a predicate is something like its cognitive significance.) There is a certain amount of overlap between an X definition and a theory of X, but the distinction between them is important. For example, a theory of planethood makes claims about the nature of planets—what it is for something to be a planet. A planethood definition might specify the extension of 'planet'—which things are planets. A semantic theory for planethood specifies the meanings of the sentences that contain 'planet' and its synonyms. It is also important to keep in mind the distinction between a semantic theory for X and a semantic theory for things that are X. For example, a semantic theory for vagueness specifies the meanings of sentences that contain 'vague' and its synonyms, whereas a semantic theory for things that are vague specifies the meanings of sentences that contain vague terms. A semantic theory for quantification specifies the meanings of sentences that contain 'quantifier' and its synonyms, whereas a semantic theory for things that display quantification specifies the meanings of sentences that contain quantifiers.

the set of sentences of L_1 and whose set of sentential meanings is a subset of the set of sentential meanings of L_1. A language L_1 is an *extension* of a language L_0 if and only if L_0 is a sublanguage of L_1. Although this definition of language as an abstract entity is popular among analytic philosophers, it leaves much to be desired.[25] Perhaps the most difficult issue facing proponents of this account of language is specifying the relation between languages and the humans who use them. I refer to this relation as *the actual language relation*.[26] A specification of the actual language relation explains what it is about the mental, physical, and social activities of a group of humans that makes them users of a particular language. I say nothing about what the actual language relation is or how one determines which language a group of people use. Another problem with this account of language is that any change in the syntactic or semantic features of the expressions used by a person or group of people results in a change in the language they use. Consequently, the language one uses changes almost continuously. Despite its deficiencies, I use this account of language because it simplifies discussions of language, and because it is the one that is assumed by most of those who propose semantic theories for truth. An attempt to construct a more plausible account of languages would take me too far afield.[27]

I follow most people who study languages by insisting on the distinction between types and tokens.[28] A *token* of a word or sentence is a physical entity (e.g., ink marks on a page, sound waves, pulses of light, etc.), while a *type* is an abstract entity. There might be many different

[25] See Lewis (1969), Soames (1984), Stalnaker (1987), and Davidson (1992).

[26] See Lewis (1969, 1975) and Schiffer (1993) for more on this issue. See Hawthorne (1990) and Field (1994a) for criticism.

[27] Some philosophers and linguists define languages in terms of mental or pragmatic elements instead of as an abstract syntactic and semantic structure. Mental definitions of language usually focus on the brain states of the humans that have linguistic capacities (e.g., Chomsky 1995), while those who favor pragmatic definitions of language often concentrate on the dispositions, regularities, or rules associated with the members of a linguistic practice (e.g., Sellars 1954 and Lewis 1975).

[28] See Kaplan (1973), Szabó (1999), and Truncellito (1999). See Kaplan (1990) for criticism. I take for granted that a person who comprehends a language can, in general, determine when a particular physical object counts as a token of a certain type of that language. Some philosophers find it useful to talk about purely abstract languages that have no tokens because these languages have never been used. To accommodate these views, I want to permit languages that have no tokens.

tokens of the same type. There might be many different tokens of the same type. For example, the previous two sentences are two tokens of the same sentence type. All the languages I consider have an infinite set of sentence types because they have sentential operators (e.g., 'and') and allow unlimited iterations of some term functions (e.g., 'the father of x'). I assume that every language has a finite set of expression tokens and a finite set of sentence tokens.[29]

So much for semantic theories and languages. The next topic is the expressibility relation. A theory T that belongs to one language L_0 is *expressible* in another language L_1 if and only if for every sentence q that composes T, there exists a sentence p of L_1 such that p is a translation of q.[30] This definition of 'expressible' relies on a notion of translation from one language into another. I assume that a sentence of one language is a translation of a sentence that belongs to another if they have the same or relevantly similar meanings (or contents).[31] Although I say very little about meaning and what makes two meanings relevantly similar, I assume that two sentences with the same or relevantly similar meanings have the same truth conditions. For the most part, I ignore issues related to the indeterminacy of translation, the indeterminacy of interpretation, the inscrutability of reference, and their implications for defining suitable notions of meaning and translation.[32] I do not want to give the impression that I think these issues are not worth discussing. Quite the contrary; there is so much to say about them that

[29] One physical object can count as a sentence token of two different languages (e.g., 'Kripke rang' is a sentence of both English and German), but such a physical object counts as two different tokens. If we erase the 'n' and add an 'ed' on the end of 'Kripke rang', then, although the result might be classified as the same physical object, it is a different sentence token of English and it is no longer a sentence token of German. Of course, not all physical changes to a physical object will change its status as a token. The 'Kripke rang' example is from Sawyer (1999).

[30] I use the word 'express' in two different ways. Words or sentences express concepts, while languages express sentences or theories. I define the latter in terms of translation and content. I say little about the former.

[31] For the most part, I ignore the distinction between meaning and content, but the standard way of drawing it is that a context dependent expression has the same meaning in every context, but its content differs from context to context. The distinction for sentential meaning and sentential content is analogous.

[32] See Quine (1960) and Davidson (1973, 1979). It is my view that there is room for accepting both Quine's thesis on the indeterminacy of translation given the behavioristic evidence he allows and Davidson's thesis on the indeterminacy of interpretation given the evidence base he allows, while at the same time accepting a perfectly legitimate notion of translation that does not commit one to the distinction between analytic and synthetic truths. However, I do not argue for this claim here.

I could not possibly give these issues the space they deserve and still discuss everything that is required to start coming to terms with the problems bequeathed to us by our concept of truth.[33]

The fourth aspect of my definition of internalizability on which I comment is application. Before defining it, I want to mention that, because of the prevalence of revenge paradoxes, it is common to treat semantic theories for truth as if they do not apply to languages that contain paradoxical sentences (i.e., the semantic theory would imply that such sentences both have and do not have some semantic property). However, I treat semantic theories as if they apply by default to every language, and I treat restrictions as explicit parts of a semantic theory—I often use the locution 'version of a semantic theory' to distinguish between semantic theories that differ only in the way they are restricted.[34] For example, Tarski's truth definition has come to be used as the standard semantic theory for first-order classical languages that do not contain their own truth predicates. However, one version of this semantic theory applies to first-order classical languages that do contain their own truth predicates. Tarski showed that if this semantic theory applies to certain languages that contain their own truth predicates, one can derive a contradiction. That is, the semantic theory implies that some sentences of these languages are both true and not true (the sentences for which this occurs are like liar sentences). According to my convention, the version of Tarski's theory that applies only to languages that do not contain their own truth predicates is one semantic theory and the version that applies to languages that do contain their own truth predicates is another. The former is consistent and the

[33] Given that I do not define 'sentential meaning', one can think of my definitions of 'language' and 'expression' as definition schemata—as the forms of definitions of 'language' and 'expression'. It does not matter for my purposes how one explains sentential meanings (e.g., in terms of sets of possible worlds, structured propositions, inferential roles, causal relations, nomic relations, etc.). One might worry that translation depends on one's choice of semantic theory. Although I disagree, a defender of this view can still accept my arguments by relativizing translation to a semantic theory. Of course, all the definitions that depend on translation are then relativized to a semantic theory as well, but that does not affect the cogency of the arguments.

[34] My approach is considerably more liberal than Davidson's. For Davidson's views on what it is for a certain theory to apply to a given language, see Davidson (1967, 1973).

latter is not. This convention of treating the restrictions as explicit additions to the theory is intended to avoid equivocations.

A *restriction* for a semantic theory T is a claim that T does not provide the meanings for the sentences of certain languages or that T does not provide the meanings of certain sentences of certain languages.[35] An *unrestricted semantic theory* is one that has no restrictions, and a *restricted semantic theory* is the conjunction of an unrestricted semantic theory and its restrictions. A semantic theory *applies to a language L* if and only if the semantic theory does not contain a restriction specifying that it does not provide the meanings for sentences of L. A semantic theory *applies to a sentence of L* if and only if it is not restricted from doing so. I say that the *scope* of a semantic theory T is the set of sentences to which T applies.[36] I assume that a semantic theory has no consequences for sentences outside its scope.

I employ a deductive account of theory application. Assume that T is a semantic theory for X and that S is the set of all the sentences in T's scope. For each member of S, an assignment follows from the union of the set of sentences that constitute T and a set of additional claims.[37] An *assignment* is a specification of the meaning of a sentence in the scope of the semantic theory in question. I have emphasized that the assignments of a semantic theory need not have the

[35] There are at least two ways to interpret the restriction claim. One is that it is a stipulation for how the unrestricted semantic theory *should be* used (i.e., one should not use it to assignments for such and such sentences); the other is that it is a claim about how the semantic theory *is* used (i.e., such and such sentences are not within the scope of this theory). On the first version, if one were to go ahead and use the restricted theory (the conjunction of the unrestricted theory and the restriction clause) in a way that is incompatible with its restrictions, then one would be guilty of some sort of pragmatic contradiction (perhaps one espouses a self-refuting theory). On the second version, if one were to do that, then the restricted theory would be false since the restriction clause would incorrectly describe the scope of the restricted theory. I prefer the second interpretation of the restriction clause because it avoids messy issues having to do with self-refutation.

[36] The assumption that this collection is a set plays no role in my presentation or arguments other than ease of exposition.

[37] The set of additional claims might involve syntactic, semantic, or pragmatic information about the sentences in S (e.g., that certain sentence is declarative, that a certain name names a particular object, or that a certain sentence token has been used to make an assertion).

form: ⟨p⟩ means that q; instead, most semantic theories assign truth-values to sentences under certain conditions.

To determine the assignments of a given semantic theory, one requires a theory of logical consequence for a set of sentences, which specifies the sentences that follow from each subset of that set. I assume that each language will require its own theory of logical consequence. When comparing sentences from different languages, I invoke the notion of translation. However, I often ignore this complication and write as if a sentence of one language is a logical consequence of a set of sentences that belong to another. A semantic theory T together with a set of auxiliary claims (e.g., claims about the syntactic structure of the sentences in the scope of T) entail, on the theory of logical consequence, an assignment for each member of the scope of T. For example, if T is a semantic theory for truth and T applies to a language L, then T provides the meanings for the sentences of L that are members of the scope of T. An assignment for each member of L that is in the scope of T follows from T.

Semantic theories can be restricted so that they apply to only a fragment of a language. Assume that a semantic theory T for X specifies the meanings of the sentences of a language L that expresses X. The scope of T for L is the set of sentences of L for which the theory delivers an assignment. Yet the scope of T for L need not be all the sentences of L that express X. Indeed, T might apply to only a proper subset of the set of sentences that express X. I call the sentences of a language that express a certain concept X the *X-sentences* of the language, and a language that contains an X-sentence I call an *X-language*. I assume that a semantic theory for X is restricted by default to X-languages and to the X-sentences of X-languages. For example, assume that a theorist is constructing a version of Tarski's semantic theory for truth that is supposed to apply to English, but she does not want the theory to be inconsistent; the theorist should restrict the scope of the theory so that it does not provide an assignment for sentences like the liar sentence that cause trouble for Tarski's semantic theory. That is, the theorist should restrict the scope of the semantic theory so that it does not apply to all the truth-sentences of English.

Tarski uses the terms *object language* and *metalanguage* to distinguish between the language to which his truth definition is intended to apply and the language in which the truth definition is formulated.[38] According to this usage, the metalanguage is expressively richer than the object language. Tarski does not say exactly what he means by 'expressively richer', but the idea is that some concepts can be expressed in the metalanguage that cannot be expressed in the object language. It turns out to be rather difficult to characterize what 'expressively richer' means in the case of Tarski's truth definition.[39]

I would like to introduce two terms to mark a distinction that is similar to the one Tarski draws. I use 'employed language' and 'target language' instead of Tarski's terms because his terms have connotations associated with them that I want to avoid. According to my usage, the *employed language* is the language in which a semantic theory is formulated, and a *target language* is a language to which that semantic theory applies. I do not require that the employed language be expressively richer than the target language(s). Indeed, if a semantic theory is internalizable for some language, then the theory's employed language and its target language might be the same language.

There is a sense in which a semantic theory for X that applies to a language L *should* provide an assignment for every X-sentence of L. It will be helpful to have a term for semantic theories that satisfy this demand.

[38] Tarski (1933).

[39] See DeVidi and Solomon (1999) for discussion of what Tarski meant by 'expressively richer'.

> A semantic theory T for X is *descriptively complete for L* if and only if T provides an assignment for every X-sentence of L.
>
> A semantic theory T for X is *descriptively incomplete for L* if and only if it is not the case that T is descriptively complete for L.

The notion of descriptive completeness plays an important role in the definition of internalizability and in the arguments of Chapter Two.

Now that I have discussed the components of the definition of internalizability and I have introduced some new terminology, I can provide a more economical definition of it and related terms.

> A semantic theory T for X is *internal for L* if and only if T is expressible in L and T is descriptively complete for L.
>
> A semantic theory T for X is *external for L* if and only if it is not the case that T is internal for L.
>
> A semantic theory T for X is *internalizable for L* if and only if there exists an extension L′ of L such that T is internal for L′.
>
> A semantic theory T for X is *essentially external for L* if and only if it is not the case that T is internalizable for L.[40]

I claim that these distinctions allow one to formulate the intuition expressed by McGee and Reinhardt: for a semantic theory to successfully apply to a natural language it must be internalizable for some language.[41] Notice that the notion of descriptive completeness does much of the work in the definition of internalizability because one can always extend a given language to express a particular semantic theory. That is, if T is a semantic theory for X that applies to an X-language L, then L can be extended to a language L′ in which T can be

[40] Although these definitions hold only for semantic theories, one could define similar notions for theories of meaning. Brandom's expressive completeness requirement seems to be something like an internalizability requirement for theories of meaning. According to Brandom, if no semantic theory that meets the conditions specified by a theory of meaning T provides the right assignments to the sentences that constitute T, then T is unacceptable. See Brandom (1994).

[41] Although in the passage I quoted above, McGee seems to claim that for a semantic theory to successfully apply to a natural language, it must be internal for that language, in later writings he makes it clear that he is interested in internalizability instead of internality. See McGee (1997: 405-406).

expressed. It might even be the case that T is essentially external for L, but T is descriptively complete for L (e.g., if T is not descriptively complete for L′ because it is restricted from applying to some of the X-sentences of L′). Internalizability for L requires that T be both expressible in L′ and descriptively complete for L′.

Notice also that the definition of internalizability does not have any implications for the correctness or truth of a semantic theory. If a semantic theory T for X is descriptively complete for an X-language L, but is essentially external for L, then one can construct a new theory T′ that is internalizable for L. One simply picks an extension L′ of L in which T can be expressed, and one stipulates that T′ agrees with T on the sentences of L and T′ implies that every sentence of L′ that is not a sentence of L is necessarily false (any other semantic concept expressible in L′ would work just as well). T′ is internalizable for L because there exists an extension L′ of L such that T′ is expressible in L′ and T′ is descriptively complete for L′. Of course, T′ is certainly false because for every X-sentence p of L′ that does not belong to L, T′ implies that both p and ⌈~p⌉ are necessarily false. The moral is that internalizability is relatively easy to achieve if one is willing to sacrifice correctness. The difficult task, when it comes to semantic theories for truth, is a theory that is both correct and internalizable for some language (by ‘correct’, I mean that it is correct not only for the truth-sentences of L, but for all the sentences in its scope); even more difficult is the task of constructing a semantic theory for truth that is correct and internalizable for a natural language.

1.3.2 Example: Kripke's Semantic Theory

In this subsection I apply the conceptual machinery I developed above to Kripke's semantic theory for truth. In 1975, Kripke published what is probably the most influential essay on truth since Tarski's truth definition of 1933. Kripke describes a procedure by which a truth predicate ('true-in-L') is introduced into a first-order language L.[42] One important feature of Kripke's approach is that the truth predicate is a partial predicate. That is, the truth predicate's extension and anti-extension are not jointly exhaustive. Kripke describes a procedure by which all the sentences of L that do not contain 'true-in-L' are placed in either the extension or the anti-extension of 'true-in-L', while some (but not all) of the sentences that contain 'true-in-L' are assigned to either the extension or anti-extension of 'true-in-L'. One way of thinking about Kripke's theory is that it assigns truth-values to some of the sentences of L under certain conditions. Those that are assigned truth are placed in the extension of 'true-in-L' and those that are assigned falsity are placed in the anti-extension of 'true-in-L'.[43] Sentences without truth-values are often called *truth-value gaps*, or just *gaps*. Kripke is quite explicit about his preferred interpretation of gaps: sentences assigned gaps do not express propositions, but they need not be meaningless. According to Kripke, a sentence is meaningful if it has a truth-value in some circumstances.[44] Thus, as long as there are some circumstances in which a sentence expresses a

[42] Kripke (1975). See Yablo (1982, 1985, 2003), Feferman (1982), Burgess (1986), Fitting (1986), McGee (1989, 1991, 2000), Gupta and Belnap (1993), Halbach (1997), Soames (1999), Visser (2001), Blamey (2002), Field (2002, 2003a, 2003b, forthcoming b), and Maudlin (2004) for discussion of Kripke's theory.

[43] As we will see, it is important to distinguish between the concept of truth that is expressed by L's predicate 'true-in-L' and the concept of truth employed by the semantic theory. I use the term 'truth-value' to refer to the assignments of the semantic theory. In this informal presentation, I sacrifice rigor for accessibility.

[44] Kripke (1975: 699). See Blamey (2002), Maudlin (2004), and Appendix B for discussion of this aspect of Kripke's theory.

proposition (i.e., has a truth-value), it is meaningful. He also stresses that the employed language need not have gaps.[45]

Kripke shows that his procedure for placing sentences that contain 'true-in-L' in either the extension or the anti-extension of 'true-in-L' eventually reaches a point where no more sentences are placed in either one. A language with a truth predicate that has this feature is called a *fixed point*. This procedure allows Kripke to define several important semantic notions (e.g., groundedness, paradoxicality, etc.) and a whole system of different fixed points and valuation schemes.[46]

In order to illustrate how the concepts defined above apply to Kripke's semantic theory for truth, consider an example. Let us assume that Rex is a person who decides to construct a version of Kripke's semantic theory for truth that is intended to apply to English (I use 'KT' as a name for the theory Rex constructs). The truth-sentences of English are all those English declarative sentences that contain 'true' or its synonyms.[47] Rex needs to decide on the scope of KT for English. There are five issues that should affect his decision: (i) KT is expressively restricted, (ii) KT applies only to interpreted languages, (iii) KT is a semantic theory for a language-specific concept of truth, (iv) KT might give rise to revenge paradoxes, and (v) KT might face self-refutation problems. I address each of these issues in order.

[45] Kripke (1975: 700-701 n. 18). My interpretation of Kripke's semantic theory might seem to be at odds with the one given by Field, who denies that the languages Kripke considers have truth-value gaps; see Field (2003b: 270). However, Field means something different by 'true-value gap' than most of those who work on truth and the liar paradox. The common usage is that a sentence p of a language L is a truth-value gap if p is a member of neither the extension nor the anti-extension of 'true-in-L'. On Field's usage, a sentence p of L is a truth-value gap if the sentence 'p is not true-in-L and p is not false-in-L' is in the extension of 'true-in-L'. The languages Kripke considers allow truth-value gaps in the common sense of the term, but do not allow truth-value gaps in the Fieldian sense of the term. There is a difference between a language that has truth-value gaps and a language that contains a true sentence that says that a particular sentence of that language is neither true in that language nor false in that language. If L is a language Kripke considers, then some sentences of L are neither true-in-L nor false-in-L, but one cannot utter a true-in-L sentence of L that expresses this claim.

[46] A valuation scheme is roughly a method for assigning truth-values to the logically compound sentences of L.

[47] To be precise, we can assume that the language in question is English as spoken by me at noon GMT on January 1st, 2004. Hereafter I assume that all the sentences I discuss are declarative.

First, Kripke's semantic theory is expressively restricted. An *expressively restricted* semantic theory is one that applies only to languages that lack certain expressive resources. In particular, Kripke's semantic theory cannot apply to languages that contain non-monotonic sentential operators.[48] For languages that contain non-monotonic sentential operators, Kripke's procedure never reaches a fixed point and, hence, it does not provide any assignments for the sentences of such languages. Thus, Rex must restrict the scope of KT for English so that it does not contain any sentences with non-monotonic sentential operators.[49]

Second, Kripke's semantic theory (like most semantic theories) applies only to interpreted languages. An *interpreted language* is one whose universe of discourse is a set—the *domain* of the language. In such a language, the predicates have extensions that are subsets of the domain, the n-place relation terms are assigned sets of n-tuples of objects from the domain, the singular terms pick out objects from the domain, and the quantifiers range over objects in the domain. Although for most purposes we can treat sentences of English as if they belong to an interpreted language, English seems to contain expressions and sentences that cannot be treated as if they are part of an interpreted language because they are "about" collections of things that are "bigger" than any set (e.g., some expressions and sentences of set theory). Hence, for KT to

[48] In a three-valued scheme, a sentential operator is *monotonic* if and only if for a sentence containing that sentential operator, changing a component of that sentence from a gap to a truth-value (i.e., from a gap to true or from a gap to false) never results in changing the sentence from one truth-value to the other or from a truth-value to a truth-value gap (i.e., from true to false, from false to true, from true to a gap, or from false to a gap). Intuitively, one can "fill in" the gaps in the components without changing the truth-value of the compound. For example, choice negation ($\sim$) is monotonic, while exclusion negation ($\neg$) is non-monotonic: if $\langle p\rangle$ is true, then $\langle \sim p\rangle$ is false and $\langle \neg p\rangle$ is false; if $\langle p\rangle$ is false, then $\langle \sim p\rangle$ is true and $\langle \neg p\rangle$ is true; if $\langle p\rangle$ is a gap, then $\langle \sim p\rangle$ is a gap and $\langle \neg p\rangle$ is true.

[49] Some philosophers claim that there are no non-monotonic sentential operators; see Parsons (1984), Priest (1990), Tappenden (1999), and Maudlin (2004). I do not want to pause to consider these views in detail, but I reject them for two reasons. First, I can define a non-monotonic sentential operator and use it (I did so in footnote 48, and logicians who study many-valued logics have been doing so for decades). I find it radically implausible that everyone who has ever employed a non-monotonic sentential operator was using a meaningless expression. Second, those who deny the existence of non-monotonic sentential operators do so in an attempt to avoid revenge paradoxes for their accounts of truth. I have yet to see an independent argument for the claim that there are no such things.

apply to English, Rex will have to restrict its scope to sentences that can be treated as if they are members of an interpreted language.

Third, Kripke's semantic theory is not actually a semantic theory for truth, it is a semantic theory for truth-in-L for some particular language L. The term 'true-in-L' acts just like 'true' when applied to sentences of L, but yields either false or truth-valueless sentences when applied to sentences of other languages (depending on how it is defined). Because Rex wants KT to apply to sentences of English that express truth, he can either treat the predicate 'true' of English as 'true-in-English' and assume that sentences of English that attribute truth to sentences of other languages are false or gappy, or he can restrict the scope of KT so that these sentences are outside T's scope. The former choice renders KT false, and the latter renders it descriptively incomplete for English.[50]

Fourth, Kripke's semantic theory itself employs certain concepts that give rise to paradoxes. In the introduction, I discussed the phenomenon of revenge paradoxes. A revenge paradox for KT concerns the following sentence:

(2): (2) is either false or a truth-value gap.

Of course, one could reject one of the steps of the argument used to generate the paradox, but Kripke deals with revenge paradoxes like the one associated with (2) by restricting his theory so that it does not apply to languages that have the resources to construct revenge paradoxes.[51] Rex must either reject one of the steps of the argument or restrict the scope of KT so that it does not include any sentences that contain 'gap' or the other predicates of KT that give rise to revenge

[50] There are other ways of dealing with this issue. For example, Rex can treat 'true' as if it is synonymous with 'translatable into a sentence of English that is true-in-English', he can treat it as if it is ambiguous, or he can treat it as if it is context dependent. I object to the first and second options in Appendix A and to the third option in Appendix B.

[51] Kripke (1975: 714). One can also construct revenge paradoxes for Kripke's semantic theory with the terms 'paradoxical' and 'grounded'.

paradoxes. One consequence of this restriction is that some sentences of English (those that contain both 'true' and 'gap') will fall outside KT's scope, despite the fact that they are truth-sentences of English.

It is interesting that the concept of truth Kripke's semantic theory employs gives rise to revenge paradoxes as well; thus, Rex will have to restrict the scope of his theory even more. Kripke's semantic theory is a semantic theory for weak truth predicates. A *weak truth predicate* is one for which attributions of truth have the same truth-status[52] as the target of the attribution. For example, a weak truth predicate obeys the following rules: if ⟨p⟩ is true, then ⟨⟨p⟩ is true⟩ is true, if ⟨p⟩ is false, then ⟨⟨p⟩ is true⟩ is false, and if ⟨p⟩ is a gap, then ⟨⟨p⟩ is true⟩ is a gap as well. A *strong truth predicate* is one that behaves like a weak truth predicate except that in the case where ⟨p⟩ is gap, ⟨⟨p⟩ is true⟩ is false. If a set of truth attributions (where a truth attribution is a sentence of the form ⟨⟨p⟩ is true⟩ or ⟨~ ⟨p⟩ is true⟩) contains sentences that are gaps, then the truth predicate that appears in these truth attributions is a weak truth predicate. Truth attributions that contain a strong truth predicate are not assigned gaps.

The following is an argument for the claim that the truth predicate Kripke's semantic theory employs gives rise to a revenge paradox. First, KT can be formulated in a bivalent language (i.e., one that does not have gaps). Second, some of the languages to which KT applies do have gaps. Third, some of the consequences of KT are truth attributions (e.g., ⟨⟨p⟩ is true⟩ where ⟨p⟩ is a member of the scope of KT). Fourth, if both ⟨⟨p⟩ is true⟩ and ⟨⟨q⟩ is a gap⟩ are sentences that belong to the language in which KT is formulated, then ⟨⟨q⟩ is true⟩ and ⟨⟨p⟩ is a gap⟩ belong to this language as well. A consequence of these four claims is that KT does not

[52] I use 'truth-status' in a loose way to include truth-values and any of the ways a sentence can lack a truth-value. For example, some sentences have the truth-status of being true, some the truth-status of being false, and some have the truth-status of being a gap.

employ a weak truth predicate. Assume otherwise; if ⟨p⟩ is a sentence in the scope of KT and ⟨p⟩ is a gap, then the sentence ⟨⟨p⟩ is true⟩ (which belongs to the employed language for KT) is a gap because ⟨p⟩ is a gap and the truth predicate of the target language is weak. However, we know that no sentence of the employed language for KT is a gap because the employed language is bivalent. Hence, KT does not employ a weak truth predicate. Therefore, KT employs a truth predicate that is different from the one that belongs to its target languages. Indeed, KT employs a strong truth predicate, but it applies to languages that contain only weak truth predicates.[53]

To get the revenge paradox, consider the following sentence:

(4): (4) is not strong true.[54]

An analog of the reasoning used to derive a contradiction from sentence (2) shows that sentence (4) is paradoxical as well. Kripke's method of avoiding the paradox by assigning a gap to the paradoxical sentence does not work for (4). Rex must either reject a step of the reasoning that shows (4) is a revenge paradox or restrict the scope of KT so that it does not apply to sentences that express strong truth, which includes some of KT's consequences. If he chooses the latter, KT cannot be used as a semantic theory for some of its consequences.

Fifth, Rex needs to restrict KT so that it is not self-refuting. In the introduction, I discussed the self-refutation problem. For KT, this problem concerns the following sentence:

[53] I am assuming that the only options are a weak truth predicate or a strong truth predicate. In reality there are other options. My point is that KT employs a truth predicate that is stronger than a weak truth predicate. In fairness to Kripke, he never explicitly says that his semantic theory employs the same kind of truth predicate as the one that belongs to the languages to which it applies. Kripke does explain how his semantic theory can be altered to apply to a language with a strong truth predicate (through a procedure he calls "closing off"). However, it should be obvious that the truth predicate that belongs to the closed off language is different from the one that is employed by the semantic theory for that language; see Kripke (1975: 715). See Kremer (2000) and Field (2003a, 2003b) for related points.

[54] Note that sentence (4) does not employ exclusion negation, which is not allowed in the scope of Kripke's theory because it is non-monotonic. Sentence (4) uses choice negation, which is monotonic. Of course, those familiar with Kripke's theory already know that a sentence like (4) would cause a problem for Kripke's theory; my observation is that Kripke's theory employs the strong truth predicate that occurs in (4) instead of the benign weak truth predicate the theory purports to describe.

(3): (3) is not true.

Reasoning analogous to that for sentences (1), (2), and (4) shows that (3) is paradoxical. Thus, it seems that Kripke's semantic theory should classify (3) as a gap. Sentences that are gaps are not true. Thus, according to Kripke's semantic theory, (3) is not true (because it is a gap). However, the proposition that (3) is not true is what is expressed by (3). Therefore, (3) itself is a consequence of Kripke's semantic theory, and Kripke's semantic theory implies that (3) is not true. Consequently, Kripke's semantic theory is self-refuting; hence, it is false. My view is that the self-refutation problem, like the problem posed by revenge paradoxes, should be used with caution as a criticism of semantic theories for truth.[55] I discuss both at length in Chapter Three. If the self-refutation problem is a legitimate concern, then Rex should restrict KT so that it does not have any consequences that it labels untrue.

Is KT internalizable for English? It will be instructive to answer this question in stages. For now I assume that KT is restricted so that: (i) it does not apply to sentences containing non-monotonic sentential operators, (ii) it applies only to sentences that can be treated as members of an interpreted language, and (iii) it does not apply to sentences of a language that attribute truth to sentences of other languages (or to non-sentences). I discuss restrictions for revenge problems and self-refutation problems below. Whether KT is internalizable for English depends on what one takes the theory to be. English is certainly capable of expressing KT, but because KT is expressively restricted (i.e., it applies only to languages that lack certain expressive resources—for KT, the languages to which it applies do not have non-monotonic sentential operators), it is not descriptively complete for English or any extension of English. The reason is that some sentences of English contain both a truth predicate and a non-monotonic sentential operator;

[55] For example, the self-refutation problem for KT does not pose any additional threat. If (3) contains a strong truth predicate, then the restriction that saves KT from the revenge paradox insures that it does not apply to (3). If KT contains a weak truth predicate then (3) is not a consequence of Kripke's theory.

these sentences are outside the scope of KT. Because KT is not descriptively complete for any extension of English, KT is not internalizable for English.

Given that KT is expressively restricted, it is more instructive to consider whether KT is internalizable for a language that has fewer expressive resources. That will allow me to separate the issue of expressive restriction from the issue of internalizability. Expressively restricted semantic theories (i.e., those that apply only to languages that lack certain expressive resources) cause one set of problems and essentially external semantic theories (i.e., those that are not internalizable) cause another set of problems. To separate these problems, consider English*, which is the sublanguage of English that contains no non-monotonic sentential operators. Is KT internalizable for English*? Although it is unclear whether KT requires non-monotonic sentential operators for its formulation,[56] I am willing to assume that it does not; it follows from this assumption and the claim that KT is expressible in English that KT is expressible in English*. KT is capable of providing assignments for all the truth-sentences of English* (provided we have semantic theories for all the other linguistic items that occur in them). Furthermore, KT is expressible in English* and descriptively complete for English*; hence, KT is internal for English*. Consequently, KT is internalizable for English*. However, KT implies that some of the sentences within its scope (e.g., the revenge liar—sentence (2)—discussed above) are both true and not true. Hence, insofar as KT is expressible in English* and descriptively complete for English*, KT is inconsistent.

Let us assume that Rex decides to restrict KT so that none of the revenge liar sentences of English* occur in its scope. Let *KT′* be the restricted version of KT. Although KT′ is expressible in English*, it is not descriptively complete for English* (because it does not provide

[56] Whether it does depends on which sentences count as those that compose the theory, and, as I have said, there is room for reasonable people to disagree on this issue.

assignments for all the truth-sentences of English); hence, it is not internal for English*. Because it is restricted so that it does not provide assignments for truth-sentences that contain ‘gap’, and KT′ requires a gap predicate for its formulation, there is no extension of English* for which KT′ is descriptively complete and in which KT′ is expressible. Thus, KT′ is not internalizable for English*.

It might seem odd to say that one version of Kripke’s theory (which is inconsistent) is internalizable for English* because it is common to treat a theory as if it is automatically restricted to avoid inconsistency. As I stated, I prefer to treat versions of a theory with different restrictions as different theories to avoid equivocations. One can consider the version of Kripke’s semantic theory for truth that Rex constructs for English (KT) as the unrestricted version of Kripke’s semantic theory for truth conjoined with the claim that it is restricted in certain ways; one can consider the version of Kripke’s semantic theory for truth that Rex constructs for English* (KT′) as KT conjoined with even more restrictions.

1.4 Comparisons

In section one, I considered several arguments for the intuition voiced by McGee and Reinhardt, which I hereafter call *the internalizability intuition*. I argued that none of these arguments is successful. It is my view that the attempts to characterize the intuition on which these arguments are based are lacking. In section two, I proposed an alternative conceptual structure designed to capture this intuition. In this section, I compare my account of internalizability to some of the other attempts to capture the internalizability intuition. In particular, I discuss universality, semantic closure, and semantic self-sufficiency.

1.4.1 Universality

A *universal language* is one that can express anything that can be expressed at all. Tarski is notable for claiming that natural languages are universal.[57] The claim that natural languages are universal could mean several different things. First, it could mean that there is a set (or at least a collection) of things that can be expressed in any language whatsoever and all of these things can be expressed in the language in question. Second, it could mean that although a natural language might not be able to express everything, it can be extended so that the extended language can express everything. Third, it could mean that for any thing there is to express, a natural language can be extended so that the extended language can express that thing.[58]

What is the relation between internalizability and universality? It is possible for a semantic theory T for X to be internalizable for a language L even though L is not universal in any of these senses. In other words, it is possible that although T is expressible in an extension of L and descriptively complete for that extension of L, there are some things that cannot be expressed in any extension of L. On the other hand, even if L is a universal language, a semantic theory for X need not be internalizable for L because internalizability requires not just expressibility but descriptive completeness as well.

[57] Tarski (1933: 164). See also Herzberger (1970b), Kripke (1975), Martin (1976), Simmons (1993), and Gupta (1997).

[58] Keith Simmons discusses the concept of semantic universality, where a language is *semantically universal* if and only if it can express any semantic concept. See Simmons (1993).

1.4.2 SEMANTIC CLOSURE

‘Semantic closure’ is another term used to classify the expressive capacity of a language. Tarski originally coined the term; according to him, a language is *semantically closed* if and only if it contains names for its expressions, it contains its own truth predicate, and one can assert all the sentences that determine the proper usage of the truth predicate in the language.[59] This term is widely misinterpreted. Tarski presented a truth definition for certain artificial languages and he claimed that no such thing was possible for natural languages, in part because natural languages are semantically closed.[60] Some have taken this claim to imply that a language is semantically closed if and only if it contains its own truth predicate (or its own semantic predicates in general).[61] On the contrary, a language can contain its own truth predicate and fail to be semantically closed if its capacity for representing its syntax is restricted.[62] Others have assumed that semantically closed languages are those whose semantic theories can be expressed in them.[63] However, a language can be semantically closed without being able to express its semantic theory (e.g., if the semantic theory for some concepts of the language other than truth cannot be expressed in the language itself).[64]

It is difficult to determine the relation between internalizability and semantic closure because the phrase ‘all the sentences that determine the proper use of the truth predicate in the

[59] Tarski (1944: 348).

[60] Tarski (1944).

[61] See McCarthy (1985).

[62] See Gupta (1982).

[63] Leitgeb (2001).

[64] Sweet proposes a semantic theory with the property of local semantic closure. He presents a way of specifying a hierarchy of neighborhoods for each sentence of a language. A language is *locally semantically closed* if and only if every sentence of it is contained in a neighborhood that is part of a hierarchy of neighborhoods such that every neighborhood is contained in one that contains a truth concept for it. I hope it is clear by now that this is a strange use of ‘semantic closure’. The language in question can represent its syntax, can express all the sentences that govern the use of ‘true’, and contains its own truth predicate(s). However, it does not contain a truth predicate that applies to all the true sentences of the language. See Sweet (1999).

language' is imprecise. If we assume that the sentences of a correct semantic theory for truth that is descriptively complete for L determine the proper use of the truth predicate in L, then a semantically closed language is one for which there exists a semantic theory for truth that is internal for that language.[65] However, the converse does not hold because a semantic theory for truth could be internal for a language that is not semantically closed (e.g., if the language did not have names for all its expressions). Moreover, one could certainly specify a semantic theory for truth that is internal*izable* for a language that is not semantically closed. To sum up, if the sentences of a semantic theory for truth are all those that determine the proper use of the truth predicate in a language, then there exist semantic theories for truth that are internalizable for semantically closed languages. Furthermore, requiring that a semantic theory for truth be internalizable for its target languages does not require those target languages to be semantically closed.[66]

1.4.3 Semantic Self-Sufficiency

Anil Gupta and Nuel Belnap coined the term 'semantic self-sufficiency' to capture the intuition that natural languages can express their own semantic theories.[67] A language is *semantically self-sufficient* if and only if a correct total semantic theory for that language is expressible in it.[68]

[65] See Yablo (2003) for a different interpretation of semantic closure according to which the T-sentences for each sentence of the language are the sentences that determine the proper use of the truth predicate in that language.

[66] Herzberger distinguishes between three different types of expressive capacity. Although he calls them "degrees of semantic closure," it should be obvious that none of these is properly called "semantic closure." A language is *atomically closed* if and only if it contains the means for recording the truth-value of each of its own sentences. A language is *molecularly closed* if and only if it contains the means for expressing all singular consequences of its semantic theory. A language is *generally closed* if and only if it contains the means for expressing the whole of its semantic theory. See Herzberger (1970b).

[67] Gupta and Belnap (1993: 257). See also Gupta (1997). Note that Gupta does not support the internalizability requirement and argues against the claim that natural languages are semantically self-sufficient. Gupta and Belnap's notion of semantic self-sufficiency and Herberger's notion of general closure seem to be the same. Fitch uses 'universal' to mean what Gupta and Belnap mean by 'semantically self-sufficient'; see Fitch (1968) and Martin (1976).

There are some interesting relations between internalizability and semantic self-sufficiency. First, a semantic theory for X can be internalizable for L even though L is not semantically self-sufficient and L has no semantically self-sufficient extensions. Assume that T is a correct semantic theory for truth, that T is internalizable for a language L, and that L expresses a concept, Y, for which there exists no semantic theory that is internalizable for L. L is not semantically self-sufficient. Let L′ be an extension of L such that T is expressible in L′ and T is descriptively complete for L′. Of course, as long as L contains a totally defined, two-place sentential connective, some truth-sentences of L will express Y. It might seem as if T will not be able to provide assignments for the truth-sentences of L that express Y unless the semantic theory for Y is expressible in L′. However, some reflection shows that this impression is false. Let U be a correct semantic theory for Y. Assume that we extend L′ to a new language L″ in which U is expressible. Since U is essentially external for L, U will not be descriptively complete for L″. However, a speaker of L″ can use both T and U to provide assignments for the truth-sentences of L′ that express Y. Thus, T is both expressible in L′ and descriptively complete for L′ despite the fact that neither L′ nor any extension of L′ is semantically self-sufficient. Hence, T is internalizable for L, and no extension of L is semantically self-sufficient.

If Y is *inexplicable* in the sense that no semantic theory for Y is descriptively complete for any Y-language, then T is essentially external for L (as is every semantic theory for a concept expressible in L). In that sense, demanding that semantic theories that apply to natural languages should be internalizable for those natural languages does place demands on the intelligibility of

[68] A *total semantic theory* is one that applies to every sentence of a language, while a *partial semantic theory* is one that is not total. For example, a semantic theory for truth is a partial semantic theory because it does not apply to those sentences that do not express the concept of truth. I assume that a total semantic theory for a language is descriptively complete for that language.

the other elements of natural languages. However, it does not imply that natural languages have extensions that are semantically self-sufficient.

1.5 Conclusion

The main point of this chapter is to define internalizability for semantic theories. According to my definition, a semantic theory T that purports to specify the meanings of sentences that express a concept X is *internalizable for a language L* if and only if there exists an extension of L such that all the sentences that compose T can be translated into sentences that belong to the extension of L and T specifies the meaning of all the sentences of the extension of L that express X. I discussed the various parts of this definition and provided an example of a semantic theory (Kripke's) that is not internalizable for a particular sublanguage of English (so long as the theory is consistent). Finally, I contrasted internalizability with universality, semantic closure, and semantic self-sufficiency. I argued that it is possible that a semantic theory for X is internalizable for a language even though none of the extensions of that language is universal, semantically closed, or semantically self-sufficient.

I want to close by saying that one of my goals has been to shift the emphasis away from the expressive properties of languages and toward the relations between semantic theories and languages. One of the lessons readers should take away from this dissertation is that the debate should be conducted in terms of the relations between theories and languages, not in terms of the properties of languages.

In the next chapter, I provide my reasons for claiming that semantic theories that apply to natural languages should be internalizable for those languages. In Chapter Three, I explain why it is so difficult to construct a semantic theory for truth that is internalizable for a natural language. It turns out that the problem is the concept of truth, not the expressive powers of natural languages.

2.0 INTERNALIZABILITY REQUIREMENTS

2.1 INTRODUCTION

In Chapter One, I introduced the resources required to distinguish, for a given language L, semantic theories that are internalizable for L from those that are essentially external for L. If there is an extension of L such that the semantic theory in question is expressible in that extension and descriptively complete for that extension, then that semantic theory is *internalizable for L*; otherwise it is *essentially external for L*.

Recall that McGee and Reinhardt express the intuition that if a semantic theory does not apply to the language in which it is formulated, then that semantic theory is unsuitable for application to natural languages. In my terminology, this is an internalizability requirement: if a semantic theory T for X applies to a natural language then there exists a language L such that T is internalizable for L. It turns out that that my terminology allows one to distinguish between several internalizability requirements of different strengths, and this one is rather weak. In the first section of this chapter, I present and discuss five internalizability requirements, and I define some of the terms used to formulate them.

The rest of this chapter is devoted to defending two internalizability requirements. I provide two arguments—one for each of them. The first argument shows that if T is a semantic theory for X and there exists an X-language L such that T is essentially external for L, then there exists an X-language to which T does not successfully apply. The conclusion is that a semantic theory T for X should be internalizable for every X-language. The second argument is specific to semantic theories for truth and to natural languages. I argue that if a correct semantic theory for truth is essentially external for a natural language L, then T is not descriptively complete for L. The conclusion is that a semantic theory T for truth that applies to a natural language L should be internalizable for L. The first argument is in section three; the second is in section four. Each argument is accompanied by my replies to several objections.

2.2 Internalizability Requirements

In this section, I discuss what it is for a semantic theory to successfully apply to a language, and I present several internalizability requirements as way of explicating the internalizability intuition. What is it for a semantic theory to apply successfully to a language? The answer is sure to be something like: the semantic theory accurately describes the parts of the language that it is supposed to describe. In Chapter One, I introduced the notion of descriptive completeness to capture what it means for a semantic theory to describe what it is supposed to describe. Now I define descriptive correctness to capture the ‘accurately’ in ‘accurately describes’.

> A semantic theory T for X is *descriptively correct for a language L* if and only if T is consistent and T provides correct assignments for every member of L in its scope.

The consistency clause forbids theories that have correct assignments but contradictory consequences. I assume that a consistent theory has no contradictions as consequences (according to first-order classical logic).[1]

Next I define versions of 'descriptive completeness', 'descriptive correctness', and 'internalizability' that express properties of semantic theories instead of relations between semantic theories and languages.

> A semantic theory T for X is *descriptively complete* if and only if for every X-language L, T is descriptively complete for L.
>
> A semantic theory T for X is *descriptively correct* if and only if for every X-language L, T is descriptively correct for L.
>
> A semantic theory T for X is *internalizable* if and only if for every X-language L, T is internalizable for L.[2]

It is possible that a semantic theory for X is internalizable for some X-language and essentially external for a different X-language (for example, if T is a semantic theory for X that is expressively restricted so that no sentences that express the concept Y are within its scope and none of the sentences of T express Y, then T might be internalizable for some X-language that does not express Y, but it will not be internalizable for those X-languages that express Y and have a completely defined two-place sentential operator because such languages will have X-

[1] It is my view that self-refutation should be explained in terms of logical inconsistency. Thus, the consistency clause is intended to rule out both semantic theories that face revenge problems and those that face self-refutation problems.

[2] All the terms defined here are strong. Each one has a weak version as well and their antonyms have weak and strong versions. For example, a semantic theory for X is *weakly descriptively correct* if and only it is descriptively correct for some X-language. A semantic theory for X is *strongly descriptively incomplete* if and only if it is descriptively incomplete for every X-language. When I use one of the terms defined in the text without 'strong' or 'weak' I intend the strong version; when I use one of their antonyms, I intend the weak version (e.g., an *internalizable* semantic theory for X is one that is internalizable for every X-language, and an *essentially external* semantic theory for X is one that is essentially external for some X-language). That is not as counterintuitive as it sounds. A semantic theory for X is internalizable if and only if it is not essentially external. That is, a semantic theory for X is strongly internalizable if and only if it is not weakly essentially external. That is, a semantic theory for X is internalizable for every X-language if and only if it is not the case that it is essentially external for some X-language.

sentences that express Y).[3] That means that there are two notions of internalizability—a weak one and a strong one. If a semantic theory for X is internalizable for some X-language, then it is *weakly internalizable*; if a semantic theory for X is internalizable for every X-language, then it is *strongly internalizable*.

The following are two internalizability requirements:

> (STRONG) A semantic theory for X should be internalizable for every X-language (i.e., should be strongly internalizable).
>
> (WEAK) A semantic theory for X should be internalizable for some X-language (i.e., should be weakly internalizable).

Any semantic theory that satisfies (STRONG) satisfies (WEAK) as well. However, these requirements make no mention of natural languages. The following are three internalizability requirements that are specific to natural languages:

> (STRONG$_N$) A semantic theory for X that applies to a natural language should be internalizable for every X-language.
>
> (MODERATE$_N$) A semantic theory for X that applies to a natural language should be internalizable for that language.
>
> (WEAK$_N$) A semantic theory for X that applies to a natural language should be internalizable for some X-language.

A semantic theory that satisfies (STRONG$_N$) satisfies (MODERATE$_N$) and one that satisfies (MODERATE$_N$) satisfies (WEAK$_N$), but the converses of these claims are false.

The internalizability intuition voiced by McGee and Reinhardt I have set out to capture is that a semantic theory that does not apply to the languages in which it is formulated should not apply to a natural language. In my terminology, they suggest (WEAK$_N$). Instead of arguing for (WEAK$_N$) directly, I argue for (STRONG) in section four, and I argue for (MODERATE$_N$) as a

[3] McGee's semantic theory for truth is an example of a semantic theory that is both expressively restricted and internalizable for some language. See McGee (1991) and see McGee (1989) for an abstract. See also Maudlin (2004) and Field (2003a, 2003b) for examples. See Appendix D for discussion.

requirement on semantic theories for truth in section five. Both (STRONG) and (MODERATE$_N$) imply (WEAK$_N$). Moreover, there are good reasons for thinking that semantic theories that satisfy (WEAK$_N$) but not (MODERATE$_N$) are not descriptively correct and descriptively complete for natural languages.

I use the term ‘natural language’ throughout the rest of this chapter. At this point, I prefer to leave it at an intuitive level (i.e., natural languages are things like English, Russian, and Arabic). I discuss what I take to be the features of natural languages in section four when it is relevant to my argument. I would like to list two of my assumptions about languages (all languages, not just natural languages). First, *there are no concepts that are language-specific*. If there exist two languages, L_1 and L_2, such that L_1 can express a concept that L_2 cannot, then L_2 can be extended to a new language L_3 that can express the concept in question. Second, humans can learn to translate between languages, even if translation requires augmenting one or both of the languages involved. That is, for any two languages L_1 and L_2, L_1 can be extended to a language L_3 and L_2 can be extended to a language L_4 such that L_3 and L_4 are intertranslatable (i.e., for any sentence of L_3, there is a translation of it in L_4 and vice versa). To express this claim, I say that any two languages are *quasi-intertranslatable*.

2.3 Strong Internalizability

In this section, I argue that a semantic theory T for X that is essentially external for an X-language L is not descriptively complete for every X-language. In other words, if a semantic theory for X is descriptively complete, then it is internalizable.

2.3.1 The Argument

Assume that T is a semantic theory for X and that T is essentially external for an X-language L. I claim that T is not descriptively complete. Because T is essentially external for L, it is not the case that both T is expressible in L and T is descriptively complete for L. Hence, either T is not expressible in L or T is not descriptively complete for L. If T is not descriptively complete for L, then T is not descriptively complete (which is my conclusion); so if T is *not* descriptively complete for L, then I am done. Assume that T *is* descriptively complete for L. Hence, T is not expressible in L. Given that T is a theory, there is some language L′ in which T is expressible (e.g., L′ is the employed language for T). L and L′ are quasi-intertranslatable because all languages are quasi-intertranslatable. Hence, L can be extended to a language L″ and L′ can be extended to a language L‴ such that L″ and L‴ are intertranslatable. If T is expressible in L′, T is expressible in L‴; if T is expressible in L‴, then T is expressible in L″. Hence, T is expressible in L″. If T is expressible in L″, then T is not descriptively complete for L″ (because T is essentially external for L). Hence, T is not descriptively complete for L″. Therefore, T is not descriptively complete.

Here is an intuitive summary of the argument. T is essentially external for L, but T has to be formulated in some language or other; if we consider the result of adding to L whatever expressive resources it takes to express T, then T will be descriptively incomplete for that extended language. For example, assume that T is a semantic theory for truth that treats paradoxical sentences like the liar as truth-value gaps. Assume that if a revenge liar for gaps (e.g., ‘this sentence is either false or a gap’) is in the scope of T, then T is inconsistent (e.g., it implies are that the revenge liar is both true and not true). In order to keep T consistent, one restricts T so that it does not apply to sentences that contain ‘gap’. Although T might be descriptively complete for a language that does not contain a gap predicate, T will not be descriptively complete for languages that have both a truth predicate and a gap predicate because such a language will contain a sentence in which both a truth predicate and a gap predicate occur; such a sentence will not be in the scope of T. Hence, T does not provide assignments for all the truth-sentences of this language. Therefore, T is not descriptively complete for this language. Consequently, T is not descriptively complete.

Note that since expressively restricted semantic theories (ones that cannot apply to languages with certain expressive resources) are not internalizable, expressively restricted semantic theories are not descriptively complete either. To see that an expressively restricted semantic theory is not internalizable consider a semantic theory T for X such that sentences that express the concept Y are not within its scope. Notice that so long as none of the sentences of T express Y, T might be internalizable for some X-language. However, consider the extension L′ of an X-language L that results from adding a Y-expression to L. L′ is an X-language as well and so long as L′ contains a completely defined, two-place sentential operator (e.g., ‘and’), L′ will contain X-sentences that express Y and, hence, are outside the scope of T. Thus, T is not descriptively complete for L′. Because every extension of L′ will contain some X-sentences that express Y, T is not descriptively complete for any extension of L′. Therefore, T is not internalizable for L′. The moral is that the argument of this subsection casts doubt on expressively restricted semantic theories as well.

Why should one care about descriptive completeness? The most obvious requirement for a semantic theory for X is that it should explain the concept X. Philosophical explanation is a notoriously slippery concept but it is fairly straightforward in the case of semantic theories: a semantic theory T for X explains the concept X if and only if T assigns the right meaning to every sentence that expresses X. In my terminology, that means that a semantic theory for X should be both descriptively complete and descriptively correct. I have shown that only internalizable semantic theories can satisfy this demand; indeed, I have shown that only internalizable semantic theories can satisfy the descriptive completeness condition, regardless of whether they are descriptively correct. A semantic theory for X that is essentially external for some X-language will not be able to specify the meanings of all the sentences that express X. Hence, a semantic theory for X that is essentially external for an X-language does not explain X. In the next subsection, I reply to several objections.

2.3.2 Objections and Replies

Objection 1: We have no idea whether a semantic theory that is both descriptively complete and descriptively correct is even *possible*. Consider the case of truth. We are talking about a theory that works for *every* language that has a truth predicate. There is no good reason to require semantic theories to live up to these expectations. It hardly makes sense to criticize a semantic theory for failing to be applicable to a handful of sentences in a handful of languages.

Reply 1: This sort of "burden of proof" objection is common in the literature on truth; I address it here in some detail and refer back to it when it comes up later.[4] The obvious and intuitive view is that truth is a concept and there is something all sentences that express this concept have in common. That is the view we take on other areas of inquiry. A semantic theory for X that is consistent and provides correct assignments for all the X-sentences in all the X-languages is the ideal for a semantic theory for X. Such a semantic theory is descriptively

[4] See Gupta (1997) for a similar comment on semantic self-sufficiency and Priest (1998) for a similar argument against the law of non-contradiction.

complete and descriptively correct. Any semantic theory that fails to live up to this ideal is inadequate. Essentially external semantic theories (or *EE theories* as I sometimes call them) are not descriptively complete and descriptively correct. Thus, EE theories are inadequate.

A proponent of an EE theory (an *EE theorist*) can respond to this fact in several different ways. He could claim either that an EE semantic theory is actually a semantic theory for *X in some contexts* or that it is a semantic theory for a restricted version of X. I have no problem with either of these responses, and there is an important place for such theories. However, both leave us without a semantic theory for X and so without an explanation of X. Instead, an EE theorist might argue either that the obvious and intuitive view is wrong—that there really is nothing in common that calls for explanation—or that no matter how hard we try, the concept in question will remain inexplicable.[5] Either of these responses requires substantial argumentation. However, an EE theorist who points out that we do not know whether it is possible to achieve the ideal of descriptive completeness and descriptive correctness has not adequately responded to the problem. Of course we do not know if it can be done. If it turns out that we cannot accomplish it then there remains something we cannot explain. If an EE theory is the best we can do, then the best we can do is inadequate based on our understanding of what needs to be done. I admit that I cannot prove that a given semantic theory is descriptively complete and descriptively correct because I do not have access to all the languages to which it applies. However, I can prove that some theories are *not* descriptively complete and descriptively correct. Thus, even though we do not know whether it is possible to provide a descriptively complete and descriptively correct semantic theory for X, we do know both that EE semantic theories are not descriptively complete and descriptively correct, and that semantic theories that are not

[5] See Williams (1996) for the claim that the intuitive view of knowledge is wrong—there really is nothing in common that calls for explanation. See McGinn (2000) for the claim that the intuitive view of consciousness is wrong—no matter how hard we try, it will remain inexplicable.

descriptively complete and descriptively correct are inadequate. In conclusion, unless we already have good reason to believe that a descriptively complete and descriptively correct semantic theory for X is impossible, we have good reason to be unhappy with a semantic theory for X that is essentially external for some X-language.[6]

Objection 2: It is unreasonable to require a semantic theory for X to be descriptively complete for an X-language because any sentence that expresses X is an X-sentence. One should require a semantic theory for X to provide assignments only for some proper subset of the set of X-sentences of a given language. Perhaps it should be the set of X-attributions (an X-attribution is a sentence of the form: α is X) if X is a predicate.[7]

Reply 2: I do not expect a semantic theory for X to provide assignments for all the X-sentences of an X-language *by itself*. A semantic theory for X should work together with semantic theories for the other phenomena displayed by the language (concepts, quantifiers, names, demonstratives, sentential operators, etc.). Given this assumption, it is reasonable to expect a semantic theory for X to be able to provide assignments for all the X-sentences of an X-language when it can work together with other semantic theories that apply to that X-language. For example, if a semantic theory for necessity were unable to provide assignments for sentences that both express necessity and contain pronouns, even when working together with a satisfactory semantic theory for pronoun expressions, then that semantic theory for necessity would be inadequate. A semantic theory for X should be able to work together with other semantic theories to provide an assignment for any X-sentence whatsoever.

Objection 3: Let us assume that when we use a semantic theory for X to derive an assignment for a sentence that expresses both concept X and concept Y, we use both a semantic

[6] A different burden of proof objection is that I have not shown that it is possible to construct internalizable semantic theories. My reply should be obvious.

[7] Anil Gupta suggested this objection in conversation.

theory for X and a semantic theory for Y to do it. In order to demonstrate that a semantic theory for X is descriptively complete for an X-language L, we must have semantic theories for all the other linguistic items that occur in L. Hence, whether a semantic theory for X is descriptively complete and descriptively correct depends on the existence of semantic theories for all the other linguistic items that appear in X-languages. Achieving descriptive completeness and descriptive correctness is a team effort. Hence, a semantic theory for X can fail to be descriptively complete if it turns out to be impossible to provide a semantic theory for some other concept that appears in an X-language. Why should a semantic theory be held accountable for that?

Reply 3: I have shown that internalizability is a necessary condition for descriptive completeness. However, it is not a sufficient condition; an internalizable semantic theory of X might turn out to be descriptively incomplete. Perhaps it will turn out that there are inexplicable linguistic items. If there are, then no semantic theory will be descriptively complete. I agree that we should not fault a semantic theory for X for failing to be descriptively complete because some X-language contains an inexplicable linguistic phenomenon that is relatively unrelated to X. If such a phenomenon occurs in an X-language, then a descriptively incomplete semantic theory for X will be the best we can hope for. That is, if we have good reason to believe that some concept Y is inexplicable, then we would have good reason to think that a semantic theory for X need not be descriptively complete. However, that is not the case with EE semantic theories. An EE semantic theory for X *entails* that it is *not* descriptively complete. The theory itself is responsible for its inadequacy, not some unrelated inexplicable concept.

Moreover, the fact that descriptive completeness is a team effort cuts both ways. If one accepts an EE semantic theory for X, then one accepts that it is impossible to provide a descriptively complete semantic theory for *any* other linguistic item that occurs in an X-language.[8] Assume that T is a semantic theory for X that is essentially external for L. There exists an X-language for which T is not descriptively complete. Hence, there exists an X-sentence of this language for which T provides no assignment. Consider a semantic theory T′ for Y, where Y is expressed by this X-language. There exists a sentence of this X-language that is both an X-sentence and a Y-sentence and is outside the scope of T.[9] Because T cannot provide an assignment for this sentence, T′ cannot either. Hence, T′ is not descriptively complete for this language. Therefore, T′ is not descriptively complete. That is a serious problem. It means that if we accept an EE semantic theory for truth, then we give up on descriptive completeness for semantic theories for knowledge, for semantic theories for necessity, for semantic theories for moral obligation, for *every* other semantic theory one can imagine. It means that if we accept *even one* EE semantic theory, we effectively give up trying to explain any of our concepts. The best we would be able to do is to explain restricted versions of them. I, for one, am not willing to damn humanity to eternal ignorance.

Objection 4: When we do semantics, we make idealizations. Even when we focus on natural languages, we often study artificial languages and ignore certain aspects of natural languages (e.g., indexicals, demonstratives, intensional expressions, pronouns, indefinite descriptions, vagueness, ambiguity, empty names, interrogatives, imperatives, etc.). If we try to explain everything all at once, then it becomes difficult to make any progress. The criticism of EE theories seems to be like saying that a semantic theory for X is inadequate because one can make up some new term that the theory was not designed to handle. That hardly seems fair. If we accept the internalizability requirement (STRONG), we would no longer be able to make idealizations in semantics.

[8] Semantic theories that are internalizable for some languages would still be possible though.

[9] I assume that the language has some completely defined two-place sentential operator (e.g., ‘and’).

Reply 4: I am not arguing that we should give up idealizations in semantics. In the current state of the discipline, idealizations are important and helpful. The internalizability requirement (STRONG) does not imply that idealizations in semantics are illegitimate. The internalizability requirement does mandate that it is possible to use a version of the theory in question in the absence of idealizations. When presenting a semantic theory, one might want to ignore certain linguistic phenomena. However, the assumption is that the semantic theory should be compatible with an account of those phenomena. That is, the idealization can be dropped at a later time. If one thinks of the restrictions placed on EE semantic theories as idealizations then they are idealizations that cannot be dropped. Consider Kripke's semantic theory for truth, which employs a gap predicate. A restricted version of this theory is not applicable to sentences that contain gap predicates. One might think of this as an idealization. However, it constitutes a permanent idealization. It cannot be used in combination with a semantic theory for gaps without being inconsistent.

Furthermore, when one makes an idealization for a semantic theory for X by excluding some linguistic phenomenon, one must be in a position to say that the phenomenon in question is relatively unrelated to X. For example, we feel justified when giving a semantics for truth in excluding color predicates from the ideal languages we consider because the two are relatively independent. However, according to EE semantic theories for X, the very concepts that get excluded are intimately related to X. For example, an EE semantic theory for truth that uses gaps posits an important relation between truth predicates and gap predicates. Yet it excludes gap predicates from the languages it considers. And, because it is an EE theory, it is impossible to drop this idealization.

Objection 5: Not all EE semantic theories are like Kripke's semantic theory for truth. Some are restricted from applying to sentences that contain terms employed by the semantic theory because these terms do not appear in ordinary language and the theory does not explain them. However, that does not mean that the theory generates inconsistencies when providing assignments for these sentences. For example, consider again the semantic theory for truth that treats paradoxical sentences as gaps. One need not interpret the gap predicate as a totally defined predicate. If, instead, 'gap' itself has gaps then one could say that 'this sentence is either false or a gap' is neither true, nor false, nor a gap. It would be a gap for the 'gap' predicate, or a *'gap' gap*. This approach treats the revenge paradox for the gap approach in the same way that the gap approach treats the original liar paradox. Of course, this approach would itself face a revenge liar:

(5): (5) is either false or a gap or a 'gap' gap.

One could then posit gaps for the ''gap' gap' predicate and so on for a whole sequence of 'gap' predicates. Indeed one would generate a hierarchy of gap predicates. This semantic theory would then be applicable to any particular language no matter what expressive resources it has. Indeed, there are semantic theories for truth that adopt analogous strategies: Gupta and Belnap's semantic theory for truth and Field's semantic theory for truth.[10]

Reply 5: I agree that an uninterpreted semantic theory that is used as a semantic theory for truth can also be used to explain the concepts it employs. The objection claims that I was wrong to assume in the previous reply that all EE theories employ paradoxical concepts; they can be interpreted as employing concepts that are unexplained (not paradoxical), and the theory can be used to explain the very concepts it employs. I disagree. There is no reason to restrict a semantic theory for truth that can be used to explain one of the concepts it employs. Semantic

[10] See Gupta and Belnap (1993) and Field (2003a, 2003b).

theories that are essentially external have been restricted to avoid either revenge paradoxes or self-refutation problems. However, even if that claim is false, the essentially external semantic theories for truth that can also be used to explain a concept employed by the theory are not descriptively complete.

For simplicity, I focus on Gupta and Belnap's revision theory for truth.[11] According to the revision theory, truth is a circular concept in the sense that the definition of truth is circular (i.e., the definiendum appears in the definiens). The revision theory of truth is based on a theory of circular concepts, on which the meaning of a word that expresses a circular concept is captured by a rule of revision. A rule of revision specifies the semantic features of an expression given a hypothesis about its semantic features. For example, the rule of revision for a circular predicate F implies that if we assume F has a certain extension, then we can determine a different extension for F. The rule of revision specifies the extensions of F under different assumptions about the extensions of F. Although circular concepts do not have fixed semantic features, one can use the revision rule for a given circular concept to acquire information about its semantic features by considering its behavior during repeated applications. For example, if we begin with an arbitrary extension for F and apply the revision rule for F over and over, we generate a sequence of extensions for F. If a certain object b always ends up in the extension of F and stays there through repeated applications of the revision rule to different starting extensions, then b satisfies F. Likewise, if a certain object c always ends up outside the extension of F and stays there through repeated applications of the revision rule to different starting extensions, then c does not satisfy F. We can say that b is categorically F and that c is categorically not F. When used as a semantic theory for truth, the theory of circular concepts yields the revision theory for

[11] Ancestors of the revision theory I describe were developed by Herzberger and Gupta independently; see Herzberger (1982a, 1982b), and Gupta (1982). The theory I describe is the one expounded in Gupta and Belnap (1993). Field's theory is quite different and I discuss it in Appendix A.

truth.[12] For a given language L, the revision theory for truth specifies the set of categorical sentences (those that are categorically true or categorically false) and the set of uncategorical sentences.

The employed language for the revision theory for truth must be expressively richer than its target languages.[13] According to my terminology, the revision theory for truth is essentially external. In particular, it is not descriptively complete for languages that contain ‘categorical’. One way to interpret this limitation is that the revision theory is restricted from applying to sentences that contain ‘categorical’ because it implies that some of them (e.g., sentence (6) below) are both categorical and uncategorical:

(6): (6) is either false or not categorical.

Gupta and Belnap explicitly reject this interpretation because, according to them, ‘categorical’ is a circular concept as well.[14]

One of the exciting properties of the theory of circular concepts is that it can handle systems of interdependent circular concepts. Gupta and Belnap show that if one treats both ‘true’ and ‘categorical’ as circular concepts that are interdependent, then one can construct a semantic theory for a language containing both these terms that is descriptively complete for that language. Sentences like the categorical liar turn out to be uncategorical according to this theory. However, Gupta and Belnap claim that this is a different notion of categoricality from

[12] Note that what Herzberger, Gupta, and Belnap say about truth constitutes several theories of truth (i.e., accounts of the nature of truth) and several semantic theories for truth (i.e., theories that provide the meanings of sentences that contain truth-expressions). For example, the claim that truth is a circular concept is part of Gupta and Belnap’s theory of truth, but it is not part of their semantic theory for truth. I use ‘the revision theory for truth’ as a term for the semantic theory for truth that, together with their claims about the nature of truth (their theory of truth), constitutes what they call “the revision theory of truth”.

[13] Gupta and Belnap admit that this is the case and that the restriction is the result of the revenge paradox: “We have been concerned, for the most part, with languages whose only problematic element is the truth predicate. Even if these languages can be enriched, the strengthened versions of the paradoxes show that an adequate description of any of them requires, within our general framework, a richer metalanguage,” (Gupta and Belnap 1993: 256).

[14] Gupta and Belnap (1993: 229-235, 255-256).

the one expressed in the target language. To avoid confusion, I use 'categoricality$_0$' for the notion of categoricality employed by the revision theory for truth, and I use 'categoricality$_1$' for the notion of categoricality employed by the revision theory for truth and categoricality$_0$. Using Gupta and Belnap's strategy, one can construct a revision theory for truth and any finite number of categoricality concepts. Accordingly, sentence (6) should read: (6) is either false or not categorical$_0$. Whereas the liar sentence is not categorical$_0$, sentence (6) is not categorical$_1$. Gupta and Belnap claim that it is inappropriate to call the liar sentence true and inappropriate to call it false; likewise, it is inappropriate to call sentence (6) either categorical$_0$ or not categorical$_0$.[15]

I want to make two points about this approach. First, the uninterpreted[16] revision theory is used to construct a semantic theory for truth; the uninterpreted theory is then used to construct a semantic theory for truth and categoricality$_0$, the uninterpreted theory is then used to construct a semantic theory for truth, categoricality$_0$, and categoricality$_1$, etc. These are different theories. It is possible that the revision theory for truth is descriptively complete for a language that does not contain any of the categoricality predicates. It is possible that the revision theory for truth and categoricality$_0$ is descriptively complete for a language that contains a truth predicate and a categoricality$_0$ predicate but none of the other categoricality predicates. And so on. It is a revision theory for groups of interdependent concepts that works as a semantic theory for a language that contains some of the categoricality predicates. It might seem like we are adding components to the revision theory for truth to accommodate the new terms added to the

[15] Gupta and Belnap (1993: 255).

[16] I distinguish between uninterpreted semantic theories and interpreted semantic theories. An *uninterpreted semantic theory* is a mathematical or logical structure, which does not serve as a semantic theory by itself. An *interpreted semantic theory* is one that has been designated for some specific purpose. For example, Kripke's semantic theory is a mathematical structure that can be used for multiple different purposes. Kripke uses it as a semantic theory for truth, but McGee uses it as a semantic theory for definite truth. These are two different interpreted semantic theories that derive from the same uninterpreted semantic theory. See Kripke (1975) and McGee (1991).

language, but this impression is false. Truth, categoricality$_0$, categoricality$_1$, etc. must be explained together if a revision theory that explains them is to be descriptively complete for the language in which they appear. When we add a new categoricality predicate to the language, we must discard the theory that worked before and construct a new theory for the extended language.

Second, given a particular revision theory for truth, categoricality$_0$, categoricality$_1$, …, categoricality$_n$, one can produce a language for which it is not descriptively complete (e.g., one with a categoricality$_{n+1}$ predicate). Hence, no particular revision theory is descriptively complete. Moreover, there is no language for which a particular revision theory is internalizable. Given some comments in Gupta and Belnap (1993), the following scenario might seem plausible. We have a language L that contains a truth predicate and no categoricality predicates, and we have a revision theory T_0 for truth that is descriptively complete and descriptively correct for L. T_0 is not expressible in L because T_0 employs a categoricality$_0$ predicate, and L does not contain such a predicate. We extend L to L_0 by adding a categoricality$_0$ predicate to L. Now L_0 can express T_0, but T_0 is not descriptively complete for L_0 because T_0 does not apply to sentences that contain categoricality$_0$ predicates. We construct a new theory T_1 that is a revision theory for truth and categoricality$_0$. T_1 is descriptively complete for L_0 but is not expressible in L_0 because T_1 employs a categoricality$_1$ predicate and L_0 does not contain a categoricality$_1$ predicate. We continue in this manner extending the languages with categoricality predicates and constructing new revision theories. At no point do we reach a revision theory and a language such that that revision theory is both expressible in that language and descriptively complete for that language.[17] Therefore, no revision theory is internalizable. That is the most optimistic outlook.

[17] There is no reason to think that revision theories for transfinite sets of categoricality predicates (if one were to extend the revision theory in this way) would fare any better than those for finite sets.

However, I argue in Appendix C that this outlook is not accurate either; the revision theory does face revenge paradoxes.

To sum up: in my reply to the previous objection, I assumed that the concepts employed by an EE theory are paradoxical and that the idealization used to restrict the application of the EE theory could never be dropped. The objection currently under discussion holds that one can drop the idealization used to restrict the application of an EE theory if one treats the concepts employed by the EE theory as if the EE theory itself applies to them. My reply to this counter-objection is that no single EE theory is internalizable. Although the idealization used to restrict one EE theory can be dropped by using another EE theory to explain a concept employed by the first, the second will require its own idealization. Hence, one can produce a sequence of EE theories such that each one explains a concept employed by the one before it, but this is simply a process whereby one idealization is replaced with others. One never reaches an unrestricted theory or a point at which the idealizations have been dropped.[18]

Objection 6: The argument for strong internalizability is no better off than the expressive capacity argument discussed in section one because a semantic theory for truth is internalizable for a language L if and only if L is semantically self-sufficient. Thus, internalizability is just the property of semantic self-sufficiency.[19]

Reply 6: One of the strengths of the conceptual framework I offer is that it does not fall prey to the objections lodged against expressive capacity arguments. Having a semantic theory that is internalizable for a language L is a relational property of L, and it is weaker than any of the commonly cited expressive properties used in expressive capacity arguments (e.g., universality, semantic closure, and semantic self-sufficiency). In particular, it is possible that a

[18] This strategy results in a hierarchy of semantic theories and concepts. I argue that such theories are unacceptable in Appendix A.

[19] Gupta suggested this objection in conversation.

semantic theory is internalizable for a language L even though it is not the case that L is universal, semantically closed, or semantically self-sufficient and it is not the case that L has an extension that is universal, semantically closed, or semantically self-sufficient. I discussed each of these concepts in Chapter One.

To summarize this section: I have argued that semantic theories should be descriptively complete and descriptively correct and that if a semantic theory for X is essentially external for some X-language, then it is not both descriptively complete and descriptively correct. These two claims constitute a strong internalizability requirement (STRONG) on semantic theories (i.e., a semantic theory for X should be internalizable for every X-language).

2.4 MODERATE INTERNALIZABILITY, TRUTH, AND NATURAL LANGUAGES

The claim I defend in this section is weaker than the one found in the previous section in two respects. First, it pertains only to semantic theories for truth. Second, I focus on internalizability for natural languages instead of on strong internalizability. I defend the claim that a semantic theory *for truth* that is essentially external *for some natural language* is not both descriptively complete and descriptively correct *for that language*. My strategy is to show that if a semantic theory for truth is (i) essentially external for a language L, (ii) descriptively complete for L, (iii) descriptively correct for L, then L lacks certain expressive resources. I argue that no natural language lacks these resources. My conclusion is that a descriptively correct semantic theory for truth that is essentially external for a natural language is not descriptively complete for that natural language. This section serves as a reply to an EE theorist who responds to the considerations of the previous section by saying that although an EE theory for truth is not both descriptively complete and descriptively correct, it will be descriptively complete and descriptively correct for most natural languages. Furthermore, the argument I present shows that it is not just a handful of sentences that have to be excluded from the scope of a semantic theory for truth that is essentially external for a natural language. Thus, an EE theorist cannot even say that his theory will be descriptively complete and descriptively correct for a large fragment of a natural language.

I first want to consider why certain semantic theories are essentially external. As I made clear in section two, some semantic theories are expressively restricted. That is, they simply return no assignments when applied to certain languages or to fragments of languages that contain certain expressive resources (recall that Kripke's semantic theory for truth is one of these). However, not all expressively restricted semantic theories are essentially external (and not all essentially external semantic theories are expressively restricted). If one can formulate an expressively restricted semantic theory for X without using any of the X-sentences that are outside its scope, then it might well be internalizable for the languages to which it applies.[20] Although expressively restricted semantic theories have their restrictions "built in" (because they simply return no assignments for the sentences of the languages in question), essentially external semantic theories often have been restricted to keep them from being inconsistent or self-refuting. That is, some semantic theories that are not expressively restricted have inconsistent or self-refuting consequences for some sentences within their scope. To avoid falsity, these semantic theories are restricted so that the sentences for which they have inconsistent or self-refuting consequences are not within their scope. Given a semantic theory T for X, *T's restricted*

[20] McGee uses a variant of Kripke's semantic theory as a semantic theory for definite truth that has this property. It is both internalizable for the language in which it is formulated and expressively restricted. See McGee (1991).

set is the set of X-sentences that do not belong to the scope of T.[21] That is, the restricted set of a semantic theory for X is the set of sentences that express X and for which T provides no assignment. For example, consider Kripke's semantic theory for truth. A restricted version of it applies only to sentences that do not contain gap predicates. Hence, sentences that contain both 'true' and 'gap' are in the restricted set of this semantic theory. The theory should apply to them (because it is a semantic theory for truth and these sentences express truth) but it does not (because it has been restricted to keep it consistent).

2.4.1 The Argument

Let T be a descriptively correct semantic theory for truth that is essentially external for a natural language L. That is, for every extension L′ of L, either T is not expressible in L′ or T is not descriptively complete for L′. The issue currently under discussion is how well T can describe L. If T is not both descriptively complete for L and descriptively correct for L, then T does not describe L well. I intend to show that T is not both descriptively correct for L and descriptively complete for L.

I first argue that T's restricted set is not empty. Given that T is essentially external for L, there exists a truth-language M such that T is not descriptively complete for M (that was the main result of section three). Thus, some truth-sentences of M are in the restricted set for T. Let *R* be the set of these sentences.

Since L contains a truth predicate, some sentences of L will be truth attributions. If L has a name 'r' for one of the members of R (the sentences of M that are outside the scope of T), then there are sentences of L that express truth but cannot fall within the scope of T. Here is the

[21] It might turn out that the collection of X-sentences that are outside the scope of T do not constitute a set. This complication will not affect my argument.

reason. T provides assignments for L's truth-sentences. T cannot provide assignments for members of R. However, 'r is true' is a sentence of L and has the same truth conditions as r. Hence, if r is outside the scope of T, then 'r is true' must be as well. Hence, T is not descriptively complete for L; there are truth-sentences of L that are not within the scope of T. Given that T is descriptively correct, that T is essentially external for L, and that L contains some minimal linguistic resources (a name for a member of R), there are truth-sentences of L that are outside the scope of T.[22]

We can weaken the assumptions about L's linguistic resources considerably. If L does not contain a name for a member of R, then L might still contain a definite description that refers to a member of R. Let $\ulcorner \iota x \phi x \urcorner$ be a definite description that refers to a member of R (where $\ulcorner \phi x \urcorner$ is a formula with 'x' as its only free variable). An analogous argument shows that $\ulcorner \iota x \phi x$ is true $\urcorner$ is a truth-sentence of L, and it cannot be a member of the scope of T. What resources does L need to contain such a definite description? It is hard to say. For example, the following definite descriptions might play the role of $\ulcorner \iota x \phi x \urcorner$: 'the sentence that John uttered yesterday at 0314 GMT', 'the first complete sentence on page 132 of the lightest book in Springfield's public library', and 'the object currently at 42.99311073141°N, 87.90563965926°W, and 632.2342 ft. above sea level'. In order to determine whether L has a definite description that refers to a member of R, one would have to know which objects are potential referents of its definite descriptions.

Even worse is the fact that one can devise coding schemes that allow one to talk about the sentences of a certain language by talking about arithmetic. Gödel's method of arithmetization is probably the most famous one of these, and it allows one to use a language to talk about its own

[22] One might object that on a strong reading of 'true', 'r is true' and r do not have the same truth conditions. My reply is that the argument goes through only for semantic theories for weak truth. However, in Chapter Seven, I show that strong truth can be defined in terms of weak truth. See also Beall (2002) on this issue.

syntactic features.[23] But one could devise others that allow one to use one language to talk about the syntactic features of another language without too much trouble. So long as the natural language in question can express arithmetic, one can use it to refer to the sentences of the restricted set in question. Of course, the coding scheme need not be expressible in the target language for it to permit one to use the target language to refer in this way.

What if L has no names and no definite descriptions and cannot express arithmetic? We can still construct truth-sentences that are outside the scope of T. For example, all we need is a quantified truth attribution that ranges over a member of R. The sentence, 'every declarative sentence is either true or false' quantifies over all declarative sentences; hence, members of R are in its range. That is all we need to construct a truth-sentence of L that is outside the scope of T, assuming we accept that if $\ulcorner(\forall x)\psi(x)\urcorner$ is true, then $\ulcorner\psi(\alpha)\urcorner$ is true where α is a member of the range of the quantifier.[24] Demonstratives cause trouble for T as well. The sentence 'that is true', where 'that' refers to a member of R, is a truth-sentence of L that cannot be in the scope of T. In order to determine whether L contains such a demonstrative sentence, one would have to know which objects are in the vicinity of the users of L. Anaphoric dependents (e.g., pronouns), propositional attitude terms (e.g., believes, knows, desires, etc.), and discourse terms (e.g., says, asserts, utters, etc.) will cause problems as well (e.g., 'it is true', 'everything Herschel believes about clowns is true', and 'everything Mel said on yesterday's episode is true').

If (i) a semantic theory T for truth is essentially external for a natural language L, (ii) T is descriptively correct, and (iii) T is descriptively complete for L, then L must be extremely impoverished. The considerations above show that if someone can use L to name, refer to, quantify over, or demonstrate a member of the restricted set for T, then T is not descriptively

[23] Gödel (1931).

[24] In this example we also need to accept that ⟨p or q⟩ is true if and only if ⟨p⟩ is true or ⟨q⟩ is true.

complete for L. Worse still is the fact that if someone can use L to name, refer to, quantify over, or demonstrate a sentence (perhaps of some language other than L) that attributes truth to a member of T's restricted set, then T is not descriptively complete for L. The same problem will occur if L can be used to name, refer to, quantify over, or demonstrate a sentence (perhaps of some other language M) that is used to name, refer to, quantify over, or demonstrate a sentence (perhaps of some language other than L or M) that attributes truth to a member of T's restricted set. And so on.

Of course, if T is a semantic theory for truth that is essentially external for a natural language, then that natural language does have the capacity to name, refer to, and quantify over the sentences of T's restricted set. Many natural languages have some linguistic device that allows for the construction of the name of a linguistic item by displaying that item. In English, one can use single quotes for this purpose (e.g., 'βαναυσία', 'découper', 'وهب' are words of Greek, French, and Arabic, respectively, but the names of these words—which occur in this very sentence—belong to English). Given coding schemes it is easy to say that a certain sentence of another language is true by talking about numbers. Natural languages also have the capacity to construct definite descriptions that refer to most anything imaginable, physical or abstract, observable or theoretical, actual or merely possible. In addition, natural languages have quantifiers and syntactic terms that allow them to express claims like 'every declarative sentence is either true or false'. All that is needed is a universal quantifier, a sentencehood predicate, and a truth predicate. Natural languages also have demonstratives, pronouns, propositional attitude terms, and discourse terms. These are the direct ways of naming, referring to, and quantifying over restricted sentences. There are innumerable indirect ways as well. One can say in English that a certain sentence of German is true, where that sentence of German is a truth attribution to a sentence of the restricted set in question. The same goes for naming, referring to, and quantifying over sentences of other languages that are truth attributions to restricted sentences, and for naming, referring to, and quantifying over sentences of other languages that name, refer to, or quantify over sentences of other languages that are truth attributions to restricted sentences, and so on. For example, one can say in English that a certain sentence of German is true, where that sentence of German attributes truth to a sentence of Portuguese, which attributes truth to a sentence of the restricted set in question. Given that natural languages have these resources, no semantic theory for truth that is essentially external for a particular natural language will be descriptively complete and descriptively correct for that language.

Before responding to some objections, I want to return to the version of Kripke's semantic theory for truth that is restricted so that no sentence containing a truth-value gap predicate, a paradoxicality predicate, a groundedness predicate, or any of the other predicates of Kripke's theory that lead to revenge paradoxes is within its scope. Assume that the employed language for Kripke's theory is English (or my idiolect of English at noon GMT on January 1, 2004) and that a target language for Kripke's theory is English*, which is a sublanguage of English. Assume that English* is a first-order language that contains a truth predicate, but it does not contain any non-monotonic sentential operators and it does not contain any of the above predicates of English to which Kripke's theory cannot apply. Kripke's theory is essentially external for English*. Hence, there exists a truth-language M such that M contains some truth-sentences that are in the restricted set for Kripke's semantic theory. We do not have to look far to find a language like M. Kripke's theory is not descriptively complete for English because none of the sentences of English that contain both a truth predicate and a truth-value gap predicate are in its scope. Thus, some sentences of English (the employed language for Kripke's theory) are in the restricted set for Kripke's theory. The issue now is: is Kripke's theory descriptively complete for English*? My answer is "no."

Let R be the set of sentences of English that are prohibited from being in the scope of Kripke's theory. If English* contains a name 'r' of one of these sentences (e.g., 'this sentence is either false or a gap'), then the sentence 'r is true' of English* will have the same truth conditions as the member of R that r names. Thus, although 'r is true' is a truth-sentence of English, it is not a member of T's scope. Hence, if English* contains a name of one of the sentences of R, then T is descriptively incomplete for English*. Likewise, if English* contains any singular term that refers to a member of R, then T is descriptively incomplete for English*. The same result holds if English* contains sentences that quantify over members of R or it contains a demonstrative that can be used to refer to a member of R. Moreover, if ⟨⟨p⟩ is true⟩ is a sentence of English* where ⟨p⟩ is a sentence of some other language that attributes truth to a member of R, then T is descriptively incomplete for English*.

2.4.2 Objections and Replies

Objection 1: Most of the semantic theories for truth are actually semantic theories for language-specific truth predicates—those that are true of only the sentences of a single language. Here is a representative quotation:

> The problem to be solved, then, is this: Given a first-order language *L* with a distinguished predicate ***T*** that means "true-in-*L*," and given a classical model M of the ***T***-free fragment of L, construct a systematic account of the signification of ***T*** that [1] yields a classification of the sentences of *L* into true/false/paradoxical/etc.—a classification that conforms to our ordinary intuitions and uses of 'true' and [2] yields an interpretation of the T-biconditionals that is in accord with the Signification Thesis, (Gupta and Belnap 1993: 32).

Note the use of the expression ‘true-in-L’. That is a truth predicate that is restricted to the sentences of a single language (L). I call a truth predicate restricted to the sentences of a single language, a *language-specific truth predicate* (an *LS truth predicate*). Semantic theories for LS truth predicates do not purport to be descriptively complete or internalizable for natural languages.

Reply 1: I agree that many of the semantic theories for truth that have been proposed are designed to apply only to artificial languages that contain LS truth predicates. The practice goes back to Tarski, who thought that attempts to provide semantic theories for the unrestricted truth predicates that occur in natural languages were futile.[25] Moreover, the vast majority of the artificial languages for which these semantic theories are designed are considered in isolation—the domains of these languages do not even include linguistic items from other languages.

Most of those who propose semantic theories for LS truth predicates have as a goal the description of truth predicates in natural languages. Here is Gupta again: “[The] revision theory, if it is to fulfill its own goals, has to be applicable to English,” (Gupta 1997: 442). Those theorists who propose semantic theories for LS truth predicates and suggest that these theories illuminate truth predicates of natural languages follow a strategy of idealization: first solve a difficult problem for certain ideal circumstances and then determine how to solve it in more complex cases. If the result of such a strategy is an EE theory for truth, then the strategy cannot work. I have argued that only semantic theories for truth that are internalizable for a natural language can be both descriptively correct and descriptively complete for that natural language. Those philosophers who propose semantic theories for LS concepts of truth are subject to the strong internalizability requirement; those who ignore the strong internalizability requirement but

[25] Tarski (1933).

suggest that their theory can be applied to natural languages are subject to the more moderate internalizability requirement.[26]

Objection 2: There is no reason that a semantic theory for an LS truth predicate should be internalizable for its target language. Such a theory applies to only one language; hence, if it is descriptively complete for that language then it is descriptively complete; the issue of whether it can be expressed in its target language is irrelevant. Moreover, because semantic theories for LS truth predicates can explain the truth predicates that occur in natural languages, there is no need for an internalizability requirement when it comes to truth.

Reply 2: I make three points in this reply: (i) semantic theories for LS truth predicates apply to multiple languages, (ii) even if we allow only LS truth predicates, the argument of section 2.4.1 still goes through, and (iii) one cannot explain a truth predicate of a natural language in terms of LS truth predicates. A semantic theory that purports to describe a single expression of a single language is useless for explaining an expression of a natural language. Any change in the language would render the semantic theory obsolete. I assume that when we talk about semantic theories for LS truth predicates, we are talking about theories that purport to provide the meanings for sentences that express the concept truth-in-L for some fixed language L. Given this assumption, the claim that an LS truth predicate for a language L must belong to L is false. Any language can be extended to include an LS truth predicate for a given language. Hence, it is not the case that a semantic theory for an LS truth predicate has only one language as its target language. Thus, its descriptive completeness does depend on its internalizability.

Moreover, even if we allow only LS truth predicates, the argument for the moderate internalizability requirement still goes through. Assume: (i) that T is a semantic theory for truth-

[26] Another "burden of proof" objection might come to mind at this point according to which I have not shown that it is possible to provide a semantic theory for an unrestricted truth predicate. My reply to such an objection is analogous to my reply to the earlier "burden of proof" objection.

in-L that applies to L and that the employed language for T is N, (ii) that t is any paradoxical sentence in N, (iii) that a language M contains a sentence s that is the translation of t, (iv) that M contains a sentence r that attributes truth-in-M to s, (v) that L contains a sentence q that is a translation of r and L contains a name for q, (vi) that L has a truth-in-L predicate and a truth-in-M predicate, and (vii) that L contains a sentence p that attributes truth-in-L to q . L contains p, and p is a truth-in-L-sentence, but p cannot be in the scope of T unless T is inconsistent. Here is the reason. If t is paradoxical, then r is paradoxical, and if r is paradoxical, then p is paradoxical (i.e., if p is in the scope of T, then T implies that p is both true-in-L and not true-in-L). Hence, if T is consistent, then p is outside the scope of T; thus, T is descriptively incomplete for L.[27]

The above argument is complex and best understood with an example. Consider again the restricted version of Kripke's semantic theory T for truth-in-L. Assume that its employed language (language N in the previous paragraph) is English, which contains a gap predicate, and assume T is restricted to avoid revenge problems. Let t be 't is either false-in-English or gappy-in-English'. That is, t is a revenge paradox for the semantic theory for truth-in-English. Assume that L is a sublanguage of English that contains a truth-in-L predicate but no gap predicate. Because L has no unrestricted truth predicate, it might seem that L cannot contain a sentence α such that α is paradoxical if t is paradoxical. If L has a name 't' for t, then L contains 't is true-in-L'. However, 't is true-in-L' is false, not paradoxical because t is not a sentence of L. Furthermore, t is not translatable into L; hence, 't is translatable into a sentence of L that is true-in-L' is false as well. However, we can still construct a sentence of L such that it is paradoxical if t is paradoxical.

[27] This argument is inspired by remarks in Mackie (1973: 252-253).

Assume that language M is a sublanguage of English that has a gap predicate, a truth-in-M predicate, and a sentence s that is a translation of t. Let s be the only sentence written on a certain blackboard, and let M contain a sentence r, which is 'the sentence on the blackboard is true-in-M'. Assume finally that L has a sentence q that is a translation of r, and that L has a name for q. Let sentence p be 'q is true-in-L'. Sentence p belongs to L, and p is paradoxical if t is paradoxical. Here is the argument:

(a) If t is true-in-English, then r is true-in-M. (Argument: If t is true-in-English, then s is true-in-M because s is a translation of t. Sentence r ('the sentence on the blackboard is true-in-M') says that s is true-in-M. Sentence s is true-in-M. Thus, r is true-in-M.)

(b) If r is true-in-M, then p is true-in-L. (Argument: If r is true-in-M, then q is true-in-L because q is a translation of r. Sentence p ('q is true-in-L') says that q is true-in-L. Thus, p is true-in-L.)

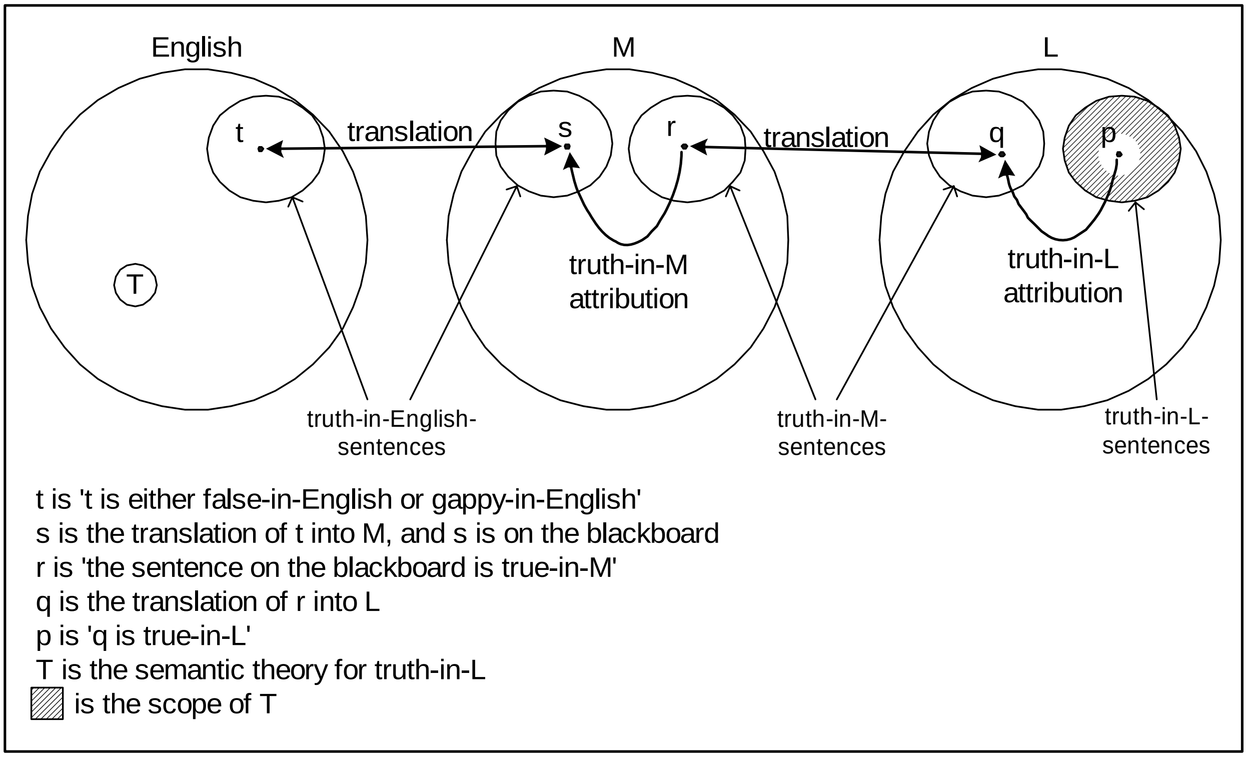
(c) If t is false-in-English or gappy-in-English, then r is false-in-M or gappy-in-M. (Argument: If t is false-in-English or gappy-in-English, then s is false-in-English or gappy-in-English because s is a translation of t.[28] Sentence r says that s is true-in-M. Sentence s is false-in-M or gappy-in-M. Thus, r is false-in-M.)

(d) If r is false-in-M or gappy-in-M, then p is false-in-L or gappy-in-L. (Argument: If r is false-in-M or gappy-in-M, then q is false-in-L or gappy-in-L because q is a translation of r. Sentence p says that q is true-in-L. Thus, p is false-in-L.)

∴ (e) If t is true-in-English iff t is false-in-English or gappy-in-English, then p is true-in-L iff p is false-in-L or gappy-in-L.

Despite the fact that L has no truth predicate for the language to which t belongs and has no sentence that is a translation of t, we have still managed to construct a sentence of L such that it is paradoxical if t is paradoxical. Thus, if p is in the scope of T, then T implies that p is both true-in-L and not true-in-L. Therefore, T is either descriptively incomplete for L or descriptively incorrect (see Figure 2.1 for a diagram of the facts used in this argument).

[28] Even if we allow multiple translations of t into M, they will all have the same truth conditions and thus will all be false or gappy.

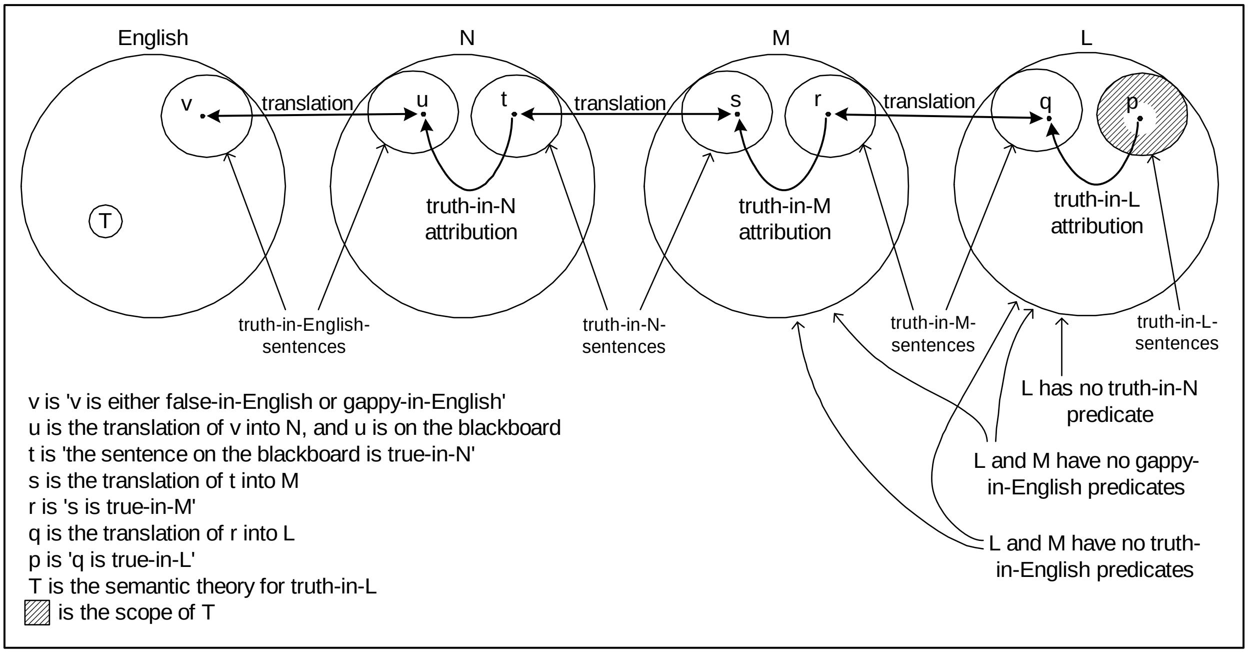


t is 't is either false-in-English or gappy-in-English'
s is the translation of t into M, and s is on the blackboard
r is 'the sentence on the blackboard is true-in-M'
q is the translation of r into L
p is 'q is true-in-L'
T is the semantic theory for truth-in-L
▨ is the scope of T

Figure 2.1 (Three-Language Situation)

Notice that, although there are many premises to this argument, all of them are quite plausible for natural languages (except the assumption that L has no unrestricted truth predicate—but the point was to show that even if L does not have an unrestricted truth predicate, T is still descriptively incomplete for L). The argument assumes that L has a LS truth predicate for a language that can express the revenge liar (t) for T. One might assume that if L had no such LS truth predicate, then T might be descriptively complete for it. On the contrary, a somewhat more complicated example shows that this assumption is false. An argument analogous to the one presented above for the situation depicted in Figure 2.2 shows that even if L has no truth-in-English predicate, no gap predicate, and no LS truth predicates for languages that have truth-in-

English predicates or gap predicates, L still contains a sentence such that it is paradoxical if the revenge paradox for T is paradoxical.



v is 'v is either false-in-English or gappy-in-English'
u is the translation of v into N, and u is on the blackboard
t is 'the sentence on the blackboard is true-in-N'
s is the translation of t into M
r is 's is true-in-M'
q is the translation of r into L
p is 'q is true-in-L'
T is the semantic theory for truth-in-L
▨ is the scope of T

Figure 2.2 (Four-Language Situation)

So far, I have made two of the three points promised in the first sentence of this reply. The third point is that truth predicates of natural languages cannot be explained in terms of LS truth predicates. That is a controversial claim and I do not have the space to defend it fully here.[29] However, I do want to present some considerations that have not received the emphasis they deserve in the literature on this issue.

The biggest problem facing LS theorists (i.e., those who claim that natural language truth predicates can be explained in terms of LS truth predicates) is that the most familiar and widely

[29] The issue of LS truth predicates arises in debates about disquotationalism, which is a version of deflationism about truth. See Leeds (1978, 1995, 1997), Williams (1986, 1999, 2002), Field (1986) (in which they are discussed but not endorsed), Resnik (1990), Quine (1992), McGee (1993), Field (1994a, 1994b), Weir (1996), Halbach (1999, 2000, 2002), and Burgess (2002). See also Appendix A.

used truth predicates in natural languages are not language-specific (e.g., in English we have ‘true’, not ‘true-in-English’, ‘true-in-Sanskrit’, ‘true-in-Klingon’, etc.). We use the English word ‘true’ to attribute truth to English sentences as well as to sentences of other languages (and to beliefs, propositions, etc.).

Two responses to this objection on behalf of LS theorists are: (i) claim that a natural language truth predicate is an *ambiguous language-specific truth predicate* (i.e., that ‘true’ is ambiguous and can be synonymous with ‘true-in-English’, ‘true-in-Sanskrit’, ‘true-in-Klingon’, etc.), or (ii) claim that a natural language truth predicate is a *translational language-specific truth predicate* (i.e., that ‘true’ is synonymous with ‘translatable into a sentence of English that is true-in-English).[30] I refer to ambiguous language-specific truth predicates as *ALS truth predicates* and I refer to translational language-specific truth predicates as *TLS truth predicates*. I address each of these proposals in order.

I have two objections to the ALS theorist. First, we do not treat our natural language truth predicates as if they are ALS truth predicates. For example, most people who understand the sentences in the following argument would say that it is valid:

(a) ‘Schnee ist weiss’ is true.

(b) ‘snow is white’ is true.

(c) ‘Schnee ist weiss’ ≠ ‘snow is white’.

∴ (d) there are at least two distinct things that are true.

If the truth predicate of English is ambiguous and takes on the meanings of different LS truth predicates in different circumstances, then this argument is invalid and suffers from an equivocation.

[30] For the latter, see Field (1986) (in which it is discussed but not endorsed), McGee (1993), Field (1994a) (in which it is endorsed as an option), Leeds (1995, 1997), and Williams (1999, 2002). For deflationist alternatives to LS truth predicates, see Field (1994a), Lance (1996), Azzouni (2001), Horwich (2001), and Brandom (2002).

Second, if our natural language truth predicates were ALS truth predicates, then we would not be able to use them properly. We routinely attribute truth blindly. That is, we assert that some sentence or set of sentences is true without knowing exactly which sentences are the targets of the attribution (e.g., everything Kripke said yesterday is true). If our truth predicate were an ALS truth predicate, then the speaker of a blind assertion of a truth attribution would have to attach a restriction or set of restrictions to the truth attribution without knowing which restrictions to attach. Thus, to use an ALS truth predicate properly, one could not use it in blind attributions. Moreover, a truth attribution to several sentences that belong to different languages would have to be either implicitly disjunctive or it would count as several attributions at once. Neither option is plausible. See Appendix A for an expanded version of this criticism.

The supporters of a TLS theory hope that by combining an account of translation with a semantic theory for a restricted truth predicate they can explain truth predicates of natural languages. Although this proposal works better than an account of an LS truth predicate alone, it is still inadequate. I offer five criticisms. First, TLS truth predicates cannot be used in blind assertions of truth attributions either. Even if one advocates a position according to which all natural languages are intertranslatable, TLS truth predicates are still more demanding than natural language truth predicates. See Appendix A for an expanded version of this criticism.

Second, as long as we treat natural languages as fixed entities, they are not intertranslatable. Let E be the idiolect of English I spoke at noon GMT on January 1, 2004 and E* the language that results from removing ‘categorical’ from E. Because the revision theory for truth can be applied to E* and is expressible in E, E is expressively richer than E*. Hence, there are sentences of E that are not translatable into E*.[31] Of course, I have assumed that any two natural languages are quasi-intertranslatable; that is, there are extensions of each that are

[31] Richard (1996), Soames (1997), and Shapiro (2003) make similar points.

intertranslatable. However, I see little hope for a quasi-translational language-specific truth predicate—one that belongs to a language L and is synonymous with ‘translatable into a sentence of an extension of L that is true-in-that-extension-of-L’. The problem is that this view tries to treat ‘L’ as both a name and a variable. Let ‘true*’ be a quasi-translational language-specific truth predicate, and let L′ be an extension of L. If ‘true*’ is synonymous with ‘translatable into a sentence of L′ that is true-in-L′’, then ‘true*’ is inadequate because there is no guarantee that L′ will be intertranslatable with the other language in question. If ‘true*’ is synonymous with ‘translatable into a sentence of some extension L′ of L that is true-in-L′’, then ‘true*’ is not an LS truth predicate at all because ‘L′’ is functioning as a variable in this expression.

Third, the hope is that a TLS truth predicate can replace a group of LS truth predicates or an ALS truth predicate, but this hope is misplaced. Assume that L and N are languages and that ⟨p⟩ is a sentence of L. Assume that I speak a different language, M. If I say in M that ⟨p⟩ is true, then that could mean that ⟨p⟩ is translatable into a sentence of M that is true-in-M or it might mean that ⟨p⟩ is translatable into a sentence of N that is true-in-N, etc. These are different truth attributions. It is possible that one is true and the other is false. If M has only ‘translatable into a sentence of M that is true-in-M’, then I cannot say in M that ⟨p⟩ is translatable into a sentence of N that is true-in-N. If I do have both truth predicates, then we run into the same problems we saw before with LS truth predicates. Thus, a single TLS truth predicate does not replace a group of LS truth predicates or an ALS truth predicate.

Fourth, many of the general principles about unrestricted truth predicates fail for TLS truth predicates. For example, a disjunction is true if and only if one of the disjuncts is true. For a TLS truth predicate, this principle becomes: a disjunction is translatable into a sentence of L that is true-in-L if and only if one of the disjuncts is translatable into a sentence of L that is true-in-L. But the principle for the TLS truth predicate is false because there are disjunctions such that one of the disjuncts but not the other is translatable into L (the left-hand side of the biconditional is false, while the right-hand side is true). Thus, what we take to be general principles for natural language truth predicates are false if natural language truth predicates are TLS truth predicates.

Finally, doing semantics requires truth predicates that apply to sentences of other languages directly. Consider a first-order language L all of whose sentential operators are monotonic. L has a single partial predicate that is interpreted as truth-in-L. In a different language, E, we formulate a version of Kripke's semantic theory T for truth-in-L. Assume that E is a classical first-order language. T provides assignments for the sentences of L, which are specifications of the truth conditions of the sentences of L. To do so, E must have a truth predicate, a falsity predicate, and a gap predicate. All of these predicates will be totally defined since E is bivalent. To avoid liar problems for E, we can assume that liar sentences of E are meaningless (this assumption is implausible as a general solution to the liar paradox but it will not cause problems in this example). Some of T's consequences are of the form ⟨⟨p⟩ is true⟩, ⟨⟨p⟩ is false⟩, ⟨⟨p⟩ is gap⟩, where ⟨p⟩ is a sentence of L. If the truth predicate of E is the LS truth predicate 'true-in-E', then one might think that T could still provide assignments for sentences of L by saying: ⟨⟨p⟩ is translatable into a sentence of E that is true-in-E⟩, etc. However, no sentence with a partial predicate can have the same content as a sentence that contains only completely defined predicates so long as two sentences with the same content have the same truth conditions. Note that when we talk about truth conditions, we usually assume that when we characterize the conditions under which a sentence is true, we also characterize the conditions under which it is false. But for sentences with partial predicates, one must characterize the truth conditions, falsity conditions, and gap conditions. The gap conditions for a sentence with a partial predicate are part of its content. Otherwise, two predicates, F and G, that have the same extension where F is completely defined and G is partially defined will have the same content no matter how different their anti-extensions are. This is obviously false. Hence, given that E is bivalent and that the truth-sentences of L contain partial predicates, the truth-sentences of L cannot be translated into E. Thus, if the truth predicate of E is 'true-in-E' or a TLS truth predicate, then all the assignments of T for truth-sentences of L are false. In order to do semantics in E for L, E must have either an unrestricted truth predicate or a truth-in-L predicate. Therefore, a TLS truth predicate is not adequate for the needs of semantics.[32]

Objection 3: The theorist who uses a semantic theory T for truth to provide assignments for truth-sentences of L need not say in advance which sentences are in the scope of T. The sentences used in the above criticism will be exceedingly rare and if, by chance, the theorist runs across one, he or she can restrict the theory at that time. For most natural languages, even semantic theories for truth that are essentially external for natural languages will be descriptively complete.

Reply 3: Why do we (philosophers) construct semantic theories? We find that we gain an understanding of natural language phenomena if we can construct a theory that will provide the right assignments for certain classes of sentences. If we want to understand truth and we feel that providing a semantic theory for truth is a good way to achieve such understanding, then the theory should work for the sentences of arbitrary natural languages that express truth. To be successful, the theory must provide assignments not only for the sentences that we find commonly bandied about; it must provide assignments for all the sentences of the language that

[32] I am assuming (contra Field) that semantic theories actually use truth predicates to provide truth conditions for the sentences of their target languages. Field distinguishes between semantic value predicates and truth predicates in an effort to avoid the revenge problems for his theory of truth. See Field (2003a, 2003b).

express truth. That is, given a sentence type of the language in question and a context for a token of that type, the theory should return an assignment for that sentence token in that context. The philosopher who constructs a semantic theory is expected to say, in advance of its application, what linguistic phenomena it is designed to illuminate. Of course, one could say, "I do not know how this theory will help our understanding; let us use it and see what it does," but no such suggestion would be taken seriously. Theorists often present their theories as restricted explanations by saying that the theory in question is a semantic theory for X, but it works only in such and such circumstances. Such theories are important because they serve as steps toward more comprehensive ones. Again, a theorist could say, "here is a semantic theory for truth and I know that it works only in some circumstances, but I cannot say which ones; let us use it and see what it does," but, as before, no such suggestion would be taken seriously. For example, did Kripke write of his semantic theory that if one discovers that a revenge liar is in its scope, then one should restrict the semantic theory so that it does not apply to that sentence? No. He provided a restriction in advance—one that would insure that the theory would not apply to sentences that might render it false.

The reason for skepticism in cases where a theorist suggests that we should restrict the theory "on the fly" is that we think that theories should not have ad hoc restrictions. The theorist who says we should restrict a semantic theory that is essentially external for a natural language if the need arises is saying something like, "my theory works except in those cases where it does not work." A semantic theory that might be inconsistent if circumstances turn unfavorable is an inadequate semantic theory.[33] Even if one were allowed to restrict one's semantic theory "on the

[33] That is an important difference between semantic theories and truth attributions. Truth attributions can be risky, but semantic theories cannot. I discuss this issue in Appendix B.

fly", the objection would still fail given that any natural language will have sentences that force the semantic theory to be restricted.

Objection 4: The claim that a descriptively correct semantic theory T for truth that is essentially external for a natural language L is not descriptively complete for L is justified by appeal to a sentence of L (say, p) that attributes truth (directly or indirectly) to a sentence of T's restricted set (say, r). The argument depends on the claim that p cannot be in T's scope because r is outside T's scope, and p has the same truth conditions as r. However, this is not sufficient reason to exclude p from T's scope. The assumption seems to be that if T cannot assign a meaning to r, then T cannot assign a meaning to p because they have the same truth conditions and hence, the same meaning. But two sentences can have the same truth conditions without having the same meaning.

Reply 4: T has plenty of consequences other than its assignments. For example, the unrestricted version of Kripke's semantic theory T for truth implies that (2) (i.e., '(2) is either false or a gap') is both true and untrue (so long as certain assumptions hold), but neither the claim that (2) is true nor the claim that (2) is untrue is an assignment of T. T assigns meanings to sentences by assigning truth conditions to sentences. T attributes truth to sentences under certain circumstances, falsity to sentences under certain circumstances, and gaphood to sentences under certain circumstances. The pressure to restrict T need not come from a problematic assignment of meaning—it can come from some other problematic consequence of T. I have assumed that T has no consequences for sentences outside its scope (other than the claim that they are outside its scope).

In the argument of section 2.4.1, I claim that if r is outside T's scope, then p must be as well. One can appeal to aletheic deflationism to justify this claim. For a deflationist, r and p ('r is true') have the same content. Thus, if r is outside the scope of T, then 'p is true' is as well. An appeal to deflationism is certainly not the only way to defend this portion of the argument. Recall the revenge liar for gaps:

(2): (2) is either false or a gap.

If sentence (2) is in the scope of Kripke's semantic theory for truth, then that theory implies that (2) is true and that (2) is not true. The following sentence is just as paradoxical as (2):

(7): (2) is true.[34]

If a proponent of the semantic theory in question restricts the theory to keep it consistent, then he must insure that both (2) and (7) are outside its scope. The fact that truth attributions to paradoxical sentences are paradoxical is well known by those who work on the aletheic paradoxes, and this is a fact about truth, not a consequence of deflationism. Gupta, one of deflationism's harshest critics, writes: "The sentence 'The Liar Sentence of Hebrew is true' is no less perplexing to us than 'the Liar Sentence of English is true'," (Gupta and Belnap 1993: 266).[35]

In this section, I have argued that if a descriptively correct semantic theory T for truth is essentially external for a natural language L and L has the expressive resources that we take all natural languages to have, then T is not descriptively complete for L. Therefore, semantic theories that are essentially external for natural languages are inadequate for use on natural languages.

[34] Provided that 'true' in (7) is a weak truth predicate.
[35] The quotation should obviously be credited to Belnap as well. Gupta's attacks on deflationism can be found in Gupta (1993a, 1993b).

2.5 INDEFINABILITY

I have argued that a semantic theory for truth should be internalizable for every truth-language (from STRONG) and for every natural language to which it applies (MODERATE$_N$). In section 2.3.2, I admitted that these conditions on semantic theories for truth are not binding if one shows that there is no possibility of satisfying them. It might seem that we do have results that suggest an internalizable semantic theory for truth is impossible; namely, Tarski's indefinability theorem and related results. In this section, I argue that such results have been misinterpreted and that, when properly understood, they give us no reason to think that an internalizable semantic theory for truth is impossible.

Let us first review Tarski's indefinability theorem. He proved that if a language L is bivalent (i.e., every sentence of L is either true or false), L is monoaletheic (i.e., no sentence of the language is both true and false), and L has the capacity to describe its own syntax, then L does not contain a predicate that is true of all and only the true sentences of L. That is, truth-in-L is indefinable in L. Given that a semantic theory for truth-in-L contains sentences that express the concept of truth-in-L, no such language can express a semantic theory for truth-in-L. Tarski proves his theorem by reductio; he shows that if L does contain its own truth-in-L predicate and satisfies the conditions of the theorem, then it contains a paradoxical sentence (i.e., a sentence for which one can derive that it is both true-in-L and not true-in-L).[36]

One can prove similar results using revenge paradoxes. For example, one can prove that a language that is not bivalent but satisfies the other conditions of Tarski's theorem does not contain a predicate with an extension that is the set of true sentences of the language and an anti-

[36] Tarski (1933). See also McGee (1985, 1991), Gupta and Belnap (1993), Simmons (1993), Halbach (1995), Soames (1999), Ketland (2000), Field (2003a, 2003b), and Maudlin (2004).

extension that is the set of untrue sentences of the language. One can prove this theorem using the revenge paradox for gap approaches to the liar paradox (i.e., 'this sentence is either false or a truth-value gap'). Another example comes from Gupta and Belnap's semantic theory for truth (i.e., the revision theory). As I mentioned in section 2.3.2, the revision theory employs the notion of categoricality; it implies that paradoxical sentences are uncategorical. Accordingly, one can construct a revenge paradox for the revision theory using the sentence 'this sentence is either false or uncategorical'. On the revision theory, this sentence is both categorical and uncategorical. Gupta and Belnap use this result to prove the indefinability of categoricality in languages that have the capacity to construct this revenge liar. One consequence is that no language that satisfies these conditions can express the revision theory. Thus, the revision theory is essentially external so long as it is consistent.[37]

One might argue that it is impossible to construct an internalizable semantic theory for truth on the following grounds. Every semantic theory for truth faces a revenge paradox. Each revenge paradox can be used to prove an indefinability theorem. Each indefinability theorem implies that the semantic theory in question is essentially external. Therefore, we have good reason to believe that an internalizable semantic theory for truth is impossible.[38]

There are a number of places at which a supporter of the internalizability requirements can attack this line of reasoning. First, it is false that every semantic theory for truth faces a revenge paradox. I agree that most semantic theories for truth face revenge paradoxes, but not all do. A semantic theory based on an error theory of truth (i.e., all sentences with truth predicates are false) does not face a revenge paradox (of course, it faces a horrible self-refutation problem, but that is quite different—one cannot prove indefinability results with self-refutation

[37] Gupta and Belnap (1993: 229-230). I elaborate on these claims in Appendix C.

[38] See Herzberger (1970a, 1981), Parsons (1983), and McCarthy (1985), which contain remarks that suggest that these philosophers would be sympathetic to this argument.

problems). One might argue that any *plausible* semantic theory for truth faces a revenge paradox or that any semantic theory for truth that does not face a self-refutation problem faces a revenge paradox, but those are not the arguments under consideration. Moreover, in my view, there are plausible semantic theories for truth that do not face revenge paradoxes or self-refutation problems. All of them are in the inconsistency tradition. That is, they all imply that truth is an inconsistent concept.[39] It is my view that revenge paradoxes and self-refutation problems result from treating what is essentially an inconsistent concept as if it were consistent. Of course, I cannot defend that claim here.[40] My point is that inconsistency theories of truth do not face revenge paradoxes and hence, the above reasoning is unsound.

Second, the indefinability results are not as strong as they appear. Let us take a look at the indefinability results in more detail. Each one has the following form: if L is a language with properties P_1, P_2, etc., and a semantic concept τ is definable in L, then L contains a sentence s and s both has and does not have some property Q. The proof concludes by rejecting the claim that τ is definable in L.

Each indefinability result has a number of hidden premises. First, they assume that the semantic concept in question is consistent. If it is an inconsistent concept, then it should not come as a surprise that it both applies and fails to apply to certain items. Moreover, if the concept in question is inconsistent, then it is not obvious that reductios are valid forms of reasoning for sentences that express this concept. Second, each result uses certain claims about the semantic concept in question to derive the contradiction. For example, Tarski's indefinability result relies on convention T (i.e., that for each sentence $\langle p \rangle$ of the language, one

[39] See Chihara (1973, 1979, 1984), Yablo (1985, 1993a, 1993b), Priest (1987), and Eklund (2002). A number of other philosophers have made remarks that suggest they are sympathetic to the inconsistency view, including Mates (1981), Barwise and Etchemendy (1987), McGee (1991), and Tappenden (1994).

[40] I defend it in Chapter Three.

can show that ⟨p⟩ is true if and only if p). A second example is that Gupta and Belnap's indefinability result for categoricality relies on the claim that the truth predicate in question obeys the revision theory for truth. Thus, each indefinability result depends on a certain theory. Moreover, each depends on the claim that the theory in question applies to the language in question. Finally, each result depends on the claim that the theory in question is consistent. Obviously, if the revision theory for truth is inconsistent, then it should come as no surprise that one could show that it implies that the sentence 'this sentence is either false or uncategorical' is both categorical and uncategorical.

Once the hidden premises of the indefinability results are made explicit, it is obvious that they pose no threat to the internalizability requirements. Each result has the form: if such and such theory is consistent, correct, and applies to such and such languages, then these languages cannot express such and such consistent concept. However, that is considerably weaker than the original formulation and it does not support the claim that no semantic theory for truth can be internalizable. Therefore, there is no good argument from the indefinability results to the claim that an internalizable semantic theory is impossible.

2.6 Conclusion

I have presented and defended several internalizability requirements on semantic theories. The strong internalizability requirement is that a semantic theory for X should be internalizable for every X-language. My argument (presented in section two) is that a semantic theory for X should be both descriptively complete and descriptively correct, and that if a semantic theory for X is descriptively complete, then it is internalizable for every X-language. A more moderate internalizability requirement is that a semantic theory for truth that applies to a natural language should be internalizable for that natural language. In other words, if a semantic theory for truth purports to specify the meanings of the sentences of a natural language that express the concept of truth, then there should be an extension of that natural language such that all the sentences of that semantic theory are translatable into that extended natural language and that semantic theory specifies the meanings of all the sentences of that extended natural language that express the concept of truth. My argument (presented in section 2.4.1) is that a descriptively correct semantic theory for truth that applies to a natural language and is not internalizable for that natural language will not be descriptively complete for that natural language.

The consequences of the internalizability requirements are far-reaching. Very few semantic theories for truth are internalizable for natural languages.[41] Even fewer are internalizable for every X-language. The internalizability requirements imply that essentially external semantic theories for truth do not illuminate the concept of truth we find in natural languages. That result turns the internalizability requirements into an effective criticism of most semantic theories for truth. In Appendix D, I discuss the semantic theories for truth that purport to be internalizable.

This concludes my discussion of internalizability. In the remainder of this dissertation, I present my positive proposals for a theory of truth and a semantic theory for truth. On my view, truth is an inconsistent concept. In Chapter Three, I explain why the fact that truth is inconsistent makes it exceedingly difficult to construct internalizable semantic theories for truth. Chapters Four, Five and Six are devoted to the construction of a theory of inconsistent concepts. In Chapter Seven, I apply this theory of inconsistent concepts to truth and derive both a theory of

[41] The only ones that purport to be are those proposed by McGee, Simmons, Field, and Maudlin. See McGee (1991), Simmons (1993), Field (2003a, 2003b, 2003c), and Maudlin (2004). See Appendix D for my evaluation of them.

truth on which truth is an inconsistent concept and a semantic theory for truth that is internalizable and does not give rise to any revenge paradoxes or self-refutation problems.

3.0 Aletheic Vengeance

3.1 Introduction

In Chapters One and Two, I argued that a semantic theory for X should be internalizable for every X-language and, in particular, that semantic theories for truth that apply to natural languages should be internalizable for those languages. Those who are unfamiliar with the literature on the aletheic paradoxes (e.g., the liar) might be surprised to find out that this requirement is so hard to satisfy that very few philosophers who present theories of truth even try to meet it.[1] The fact is that the vast majority of semantic theories for truth are not internalizable for any language. Why is it so hard to construct an internalizable semantic theory for truth? The question has rarely been asked and has never received an adequate answer.

My explanation is, roughly, that there are compelling reasons to accept that certain principles (e.g., the truth rules) are constitutive of our concept of truth; consequently, any theory of truth has to respect these constitutive principles in order to be plausible. Any theory of truth that respects these constitutive principles has to account for certain seemingly paradoxical sentences like the liar. Any theory of truth that respects the constitutive principles and

[1] I said in Chapter One that internalizability is relatively easy to achieve if one is willing to sacrifice descriptive correctness. I am assuming that no right-minded philosopher would be willing to trade descriptive correctness for internalizability. Thus, internalizability is difficult to achieve so long as one has a penchant for theories that are not trivially false.

adequately accounts for the liar sentence has four options: (i) it faces a revenge paradox, which renders it inconsistent, (ii) it faces a self-refutation problem, which renders it false, (iii) it is restricted (in order to avoid either the revenge paradoxes or the self-refutation problems), or (iv) it has the unacceptable consequence that a large class of linguistic expressions are meaningless or incoherent (in order to avoid either the revenge paradoxes or the self-refutation problems). Theories that incorporate option (i) or (ii) might be internalizable, but they are obviously false. Theories that incorporate option (iii) are usually essentially external for every language, but it is possible to construct one that is internalizable for an expressively weak language; however, none of these theories is internalizable for a natural language. Theories that incorporate option (iv) are usually internalizable for an expressively weak language, but none of them is internalizable for a natural language; in addition, they are unacceptable because they imply that certain coherent linguistic expressions are incoherent. Thus, the revenge paradoxes and the self-refutation problems work together to insure that theories of truth that respect the constitutive principles for truth are not internalizable for natural languages. I tell that story in section two.

In section three, I argue that truth is an inconsistent concept. I provide four arguments, but I admit that none of them is conclusive. However, I claim that they do give us very good reasons for pursuing theories of truth on which truth is an inconsistent concept. The first argument is that the explanation for why it is so difficult to construct a descriptively correct internalizable semantic theory for truth (from section two) gives us good reason to think that truth is an inconsistent concept. The second argument is that when one admits that truth is an inconsistent concept, one can explain the difficulties philosophers have had in constructing acceptable theories of truth; that is, one can explain why any theory of truth that respects the constitutive principles for truth faces either revenge paradoxes or self-refutation problems (unless it is restricted or it implies that certain linguistic expressions are unintelligible). No other account of truth has been able to provide such an explanation. The third argument is that McGee's theorem gives us good reason to believe that even theories of truth that do not respect the constitutive principles of truth cannot account for the way truth predicates interact with expressions of classical logic. The final argument is that, by treating truth as an inconsistent concept, one can arrive at a semantic theory for truth that is internalizable for every language. The rest of the dissertation is devoted to the presentation of such a theory. Thus, the full justification for treating truth as an inconsistent concept is not complete until the end of the dissertation.

3.2 Why is Internalizability So Hard?

Let me begin by discussing the relation between semantic theories for truth and theories of truth. Keep in mind that I am providing an explanation for the fact that it is rather difficult to provide a descriptively correct internalizable semantic theory for truth. My explanation focuses on problems faced by theories of truth. Recall that a theory of truth is just a theory that specifies some aspect of the nature of truth, while a semantic theory for truth is a theory that assigns meanings to some collection of sentences that express the concept of truth. I have discussed revenge paradoxes and self-refutation problems as problems confronting semantic theories for truth, but theories of truth face them as well. For example, a theory of truth on which bivalence holds and on which the truth rules and the inference rules of classical logic are valid implies that the liar is both true and false. Indeed, the revenge paradoxes and self-refutation problems pose a greater threat to theories of truth than they do to semantic theories for truth. The reason is that one can construct a semantic theory for truth that is not based on a truth conditional theory of meaning. These semantic theories might assign conceptual roles, assertibility conditions, or nomological roles to sentences; consequently, these theories might not imply that the sentences within their scope have any particular truth status. However, a theory of truth will imply that the sentences within its scope have certain truth statuses. Of course, the theory of truth alone might not have these consequences, but when combined with a set of auxiliary claims, it will. For example, a theory of truth might imply that a sentence is true if and only if there is a fact to which it corresponds. Alone, that theory of truth has no implications for the truth status of 'dogs are mammals', but when combined with the auxiliary claim that it is a fact that dogs are mammals and 'dogs are mammals' corresponds to this fact, it implies that 'dogs are mammals' is true. One can formulate this point by saying that theories of truth imply truth definitions—that is, some of the consequences of a theory of truth (when combined with auxiliary hypotheses) are assignments of truth statuses to the sentences within its scope. Thus, although a semantic theory for truth might be able to avoid the revenge paradoxes and self-refutation problems by assigning non-truth-conditional meanings, a theory of truth cannot.

There are two important "based on" relations pertaining to semantic theories. A semantic theory for X is based on both a theory of meaning and a theory of X. The theory of meaning on which a given semantic theory is based determines what sort of meanings it assigns to the sentences in its scope (e.g., a semantic theory for truth that is based on an inferential role theory of meaning assigns inferential roles to the sentences within its scope, a semantic theory for truth that is based on a truth conditional theory of meaning assigns truth conditions to the sentences within its scope, etc.) The theory of X on which a semantic theory for X is based lays down necessary and sufficient conditions on the semantic theory for X (e.g., a semantic theory for truth that is based on a contextual theory of truth assigns meanings to the sentences within its scope that conform to the dictates of the theory of truth—e.g., these meanings might be functions from contexts to sets of possible worlds). If a theory of truth implies that all declarative sentences are either true or false, then the semantic theory for truth that is based on this theory of truth must assign meanings to the sentences within its scope that are consistent with this principle. Obviously, the theory of meaning and the theory of X on which a semantic theory for X is based must be consistent, at least with respect to the sentences within the scope of the semantic theory. In sum, a semantic theory T for X is *based on* a theory of meaning M if and only if the meanings T assigns to the sentences in its scope conform to the conditions M lays down for meanings; a semantic theory T for X is *based on* a theory T′ of X if and only if T is consistent with the conditions T′ lays down for semantic theories for X.

One important consequence of the fact that a semantic theory for truth is based on a theory of truth is that if the theory of truth in question is restricted to avoid revenge paradoxes or self-refutation problems, then any semantic theory for truth that is based on this theory of truth will inherit these restrictions. It turns out that if a theory of truth implies that truth is a consistent concept and it respects the truth rules, then it is either: (i) inconsistent because it faces a revenge paradox, (ii) false because it faces a self-refutation problem, (iii) restricted to avoid either a revenge paradox or a self-refutation problem, or (iv) false because it implies that certain coherent linguistic expressions are incoherent. Thus, any descriptively correct semantic theory that is based on a theory of truth that both implies that truth is a consistent concept and respects the truth rules will be restricted as well. The restrictions the semantic theory inherits from the theory of truth on which it is based prevent it from being internalizable for natural languages. The rest of this section is devoted to explaining these claims.

3.2.1 Consistent Concepts and Naïve Theories

Most philosophers working on truth and the liar paradox assume that truth is a consistent concept. In fact, this assumption is so widespread that it is almost never articulated. To many, the assumption seems so obvious that to deny it is unintelligible. However, a handful of philosophers working on truth and the liar paradox have denied that truth is a consistent concept, and the theory I advocate is a member of this tradition.[2] The first order of business for someone who makes this move is to explain what an inconsistent concept is. Those who claim that truth is an inconsistent concept often disagree about the best explanation of inconsistent concepts. When I claim that truth is an inconsistent concept, I mean that the constitutive principles for truth are inconsistent. That is, simply by employing the concept of truth, one commits oneself to following incompatible rules for using it. I discuss inconsistent concepts at length in Chapters Four, Five, Six, and Seven. Here, it is enough to understand that inconsistent concepts are overdetermined for some objects. That is, the concept both applies and fails to apply to some objects. In the case of truth, it both applies and fails to apply to paradoxical sentences like the liar.

Those theorists who assume that a concept X is a consistent concept usually offer a naïve theory of X. When one offers a naïve theory of X, one proposes several principles that describe X, and one collects these principles into a theory. Thus, a *naïve theory of X* is a collection of sentences that purport to be principles that describe X. For example, the theory composed of the sentences 'every declarative sentence is either true or false', 'a sentence is true if and only if it corresponds to some fact', and 'no sentence is both true and false', constitute a naïve theory of

[2] See Chapter Seven for discussion.

truth. Axiomatic theories are common examples of naïve theories, but a naïve theory need not be formalized in this way. Naïve theories are so common that one might assume that they are the only type of theory. This assumption and the assumption that all concepts are consistent go hand in hand.

If one accepts that a certain concept X is an inconsistent concept, then one will almost certainly want to avoid a naïve theory of X. For, if X is an inconsistent concept, then it is impossible to construct a consistent naïve theory for X. That is, if the principles governing the use of X are incompatible and a naïve theory of X is a collection of these principles, then the set of sentences that constitute the naïve theory of X is inconsistent. Given that one does not want an inconsistent theory of X, admitting that X is an inconsistent concept forces one to avoid naïve theories of X. Because of this fact, the assumption that all theories are naïve and the assumption that all concepts are consistent reinforce one another: if one admits that X is an inconsistent concept, then it seems that one will have to endorse an inconsistent theory of X, and if one wants a consistent theory of X, then it seems that one will have to treat X as a consistent concept. (Some philosophers have both assumed that truth is an inconsistent concept and proposed naïve theories of truth—such theorists are forced to claim that inconsistent theories can be acceptable. This view of truth is called *dialetheism*; I discuss it in Appendix E. It seems to me that many philosophers are reluctant to admit that a given concept is inconsistent because they assume that such an admission commits them to a version of dialetheism; the theory of inconsistent concepts I offer in the next three chapters shows that this worry is unfounded.)

3.2.2 Truth Rules and Paradoxicality

I claim that the truth rules are constitutive of the concept of truth in the sense that anyone who employs the concept of truth is committed to following them (note that this claim does not mean that everyone who employs the concept of truth actually follows them). There are three truth rules:

> The *ascending truth rule* (Asc): ⟨⟨p⟩ is true⟩ follows from ⟨p⟩.
>
> The *descending truth rule* (Desc): ⟨p⟩ follows from ⟨⟨p⟩ is true⟩.
>
> The *substitution truth rule* (Sub): two names that refer to ⟨p⟩ are intersubstitutable in ⟨⟨p⟩ is true⟩ without changing its truth-value).

I do not claim that the ascending and descending truth rules are valid for every declarative sentence or even for every declarative sentence that has a truth-value. One can easily construct counterexamples to them using indexicals, demonstratives, ambiguous expressions, anaphoric expressions, etc. My claim is that they are valid for a large class of declarative sentences, which includes sentences like the liar (i.e., sentence (λ), which is '(λ) is false'). If a person uses a linguistic expression that is not governed by these rules, then that linguistic expression does not express our concept of truth. I assume that the truth rules are valid in hypothetical contexts as well as in categorical contexts. That is, one can use the truth rules in conditional arguments. In section 3.3, I discuss the options for theories of truth that deny that the truth rules are constitutive of the concept of truth.

I assume that a theory of truth that implies that truth is governed by the truth rules also implies that natural language truth predicates are univocal, invariant, and non-circular. That is, a theory of truth that implies that natural language truth predicates are ambiguous or context-dependent, or that they express circular concepts does not respect the truth rules. These theories might posit other rules that are similar to the truth rules, but they are different in significant ways. For example, a theory of truth on which natural language truth predicates are ambiguous and can have the meaning of language-specific truth predicates (e.g., ‘true-in-English’) might imply that a truth predicate of English obeys truth rules that are restricted to sentences of English. In section 3.3, I discuss the options for theories of truth that deny that natural language truth predicates are univocal, invariant, and non-circular.

Any theory of truth that respects the truth rules has to assign some truth-status to paradoxical sentences like the liar unless it is restricted so that paradoxical sentences are outside its scope. From the truth rules, the assumption that the liar is either true or false, and the rule of conditional proof, one can derive that the liar is true if and only if it is false. In classical logic, it follows that the liar is both true and false. A consistent theory of truth that respects the truth rules and allows sentences like the liar within its scope must either reject classical logic or assign some other truth-status to the liar. In particular, it is not a viable option for a theory of truth to imply that what seem to be paradoxical sentences are either ungrammatical or meaningless. The most obvious reason is that on accepted theories of syntax and meaningfulness, these sentences count as both syntactically well formed and meaningful. However, there is another reason that this strategy for dealing with paradoxical sentences is unacceptable: whether a sentence is paradoxical can be independent of the syntactic and semantic features of the sentence. Of course, the syntactic and semantic features of a sentence can guarantee that it is paradoxical; my point is that this need not be the case for every sentence. I refer to this claim as *the riskiness thesis*, and I discuss it at length in Appendix B.

Most theories of truth that respect the truth rules and have paradoxical sentences within their scope assign these sentences a truth-status other than truth or falsity. Indeed, most such theories imply that paradoxical sentences are defective in some way and, thus, they do not have truth-values. One can think of these theories as assigning these sentences the status of truth-value gaps; I call this the *gap approach*. There are many ways of interpreting truth-value gaps, but the differences between them do not affect the considerations in this chapter.[3]

3.2.3 Revenge and Self-Refutation

In this subsection, I argue that any theory of truth that respects the truth rules and has paradoxical sentences like the liar in its scope faces either a revenge paradox or a self-refutation problem. Before presenting the argument, I want to discuss both revenge paradoxes and self-refutation problems; I discuss each in connection with the gap approach. I begin with revenge paradoxes. The most well-known revenge paradox concerns the gap approach to the liar paradox. On the gap approach, the standard liar sentence, (λ) (= '(λ) is false'), is a gap. However, this theory runs into problems when confronted with the revenge liar for the gap approach:

(λ′): (λ′) is either false or a gap.

One can argue that if (λ′) is either true, false, or a gap, then it is both true and false. The argument is that if it is either true or false, then it is both true and false (the reasoning is analogous to the reasoning in the standard liar). If (λ′) is a gap, then the second disjunct of (λ′) is true; hence, (λ′) is true. Therefore, if (λ′) is either true, false, or a gap, then it is both true and false.[4] I prefer another argument that does not rely on the rule for disjunctions. This argument is analogous to the reasoning that shows that the standard liar is paradoxical. Assume that (λ′) is true. Then '(λ′) is either false or a gap' is true, and it follows that (λ′) is either false or a gap.

[3] See Yablo (1985), Soames (1999), Blamey (2002), and Field (2003a, 2003b) for discussion.

[4] See Gupta (2000).

Assume that (λ′) is either false or a gap. Then ‘(λ′) is either false or a gap’ is true, and it follows that (λ′) is true. Thus, (λ′) is true if and only if (λ′) is either false or a gap. Therefore, (λ′) is both true and either false or a gap.

The gap approach can handle the standard liar in the sense that it does not imply that the standard liar is both true and not true. However, it cannot handle the revenge liar. It implies that the revenge liar is both true and not true (reading ‘not’ in this sentence as exclusion negation). When confronted with the revenge paradox, a gap theorist has two options: (i) restrict the theory, or (ii) deny that (λ′) is either true, false, or a gap.[5] Call the first option the *modest response* to the revenge paradoxes. On the first option, semantic theories for truth that are based on the restricted theory are not internalizable for any natural language.[6] (Argument: assume otherwise; for some language L there exists an extension L′ of L such that a semantic theory for truth that is based on the restricted theory of truth in question is both expressible in L′ and descriptively complete for L′. Some of the sentences of T contain a gap predicate. Thus, L′ contains a gap predicate. Thus, L′ contains a revenge liar for the theory of truth. However, T is restricted from applying to such sentences. Hence, T is not descriptively complete for L′.)

The second option is to stipulate that the gap predicate is itself gappy. Call this the *robust response* to the revenge paradox. Hence, it is not the case that (λ′) is either true, false, or a gap; instead, it is a ‘gap’ gap.[7] Of course, this approach will face a new revenge paradox (e.g., (λ″) = ‘(λ″) is either false, a gap, or a ‘gap’ gap’). A more sophisticated version of this option is that there is a hierarchy of gap predicates, each of which can both be used to classify a revenge

[5] A gap theorist can insist that both (λ′) and ‘(λ′) is true if and only if (λ′) is either false or a gap’ are gaps as well, but this move engenders self-refutation problems. I discuss it below.

[6] See Kripke (1975) for an example. See McGee (1991), Gupta and Belnap (1993), and Field (2003a, 2003b) for discussion.

[7] See McGee (1991) for an example.

liar and figures in a new revenge liar.[8] On this theory, none of the revenge liars poses a problem. The theory implies that (λ) is a gap, that (λ′) is a 'gap' gap, that (λ″) is a ''gap' gap' gap, etc. At no point do we reach a sentence for which a contradiction follows from the theory. I discuss such theories and offer criticism in Appendix A.

One problem with this theory is that one can define a gap predicate that does not have gaps. With this new terminology, one could construct a sentence that is very much like (λ′) in the sense that the theory implies that it is both true and either false or a gap (and this consequence is a genuine contradiction). To avoid this problem the defender of the suggestion would have to make the implausible assumption that no such vocabulary could be introduced or that it is incoherent. I discuss this move below.

I want to move on to the self-refutation problem. The standard formulation of the self-refutation problem concerns the following sentence:

(ρ): (ρ) is not true.

We can reason that, on the gap approach, (ρ) must be a gap because assuming that it is either true or false leads to contradiction. However, sentences that are gaps are not true. Thus, (ρ) is not true. Notice that our conclusion, '(ρ) is not true', is identical to (ρ). Thus, (ρ) is a consequence of the gap approach.

At this point, there are two ways the self-refutation problem is commonly pursued: (i) the strong liar reasoning, or (ii) the self-refutation reasoning. Those who take the first avenue claim that, because (ρ) is a consequence of the gap approach, the gap approach implies that (ρ) is true after all, and if the gap approach implies that (ρ) is true, then it implies that (ρ) is both true and false. Hence, the gap approach is inconsistent. The conclusion of this line of reasoning is that

[8] See Field (2003a, 2003b) for an example.

the gap approach does not solve the paradox because we can use this reasoning to arrive back at the contradiction. The reasoning in short is: (ρ) is either true, false or a gap. If (ρ) is either true or false, then it is both true and false (by standard liar reasoning). If (ρ) is a gap, then (ρ) is not true. If (ρ) is not true, then '(ρ) is not true' is true (by ascending). If '(ρ) is not true' is true, then (ρ) is true (by substitution). Therefore, if (ρ) is a gap, then (ρ) is true. Consequently, if (ρ) is either true, false, or a gap, then it is both true and false. Call this the *strong liar reasoning*.[9] This result is supposed to cast doubt on the efficacy of the gap approach.

Those who travel down the other path from the observation that the gap approach has (ρ) as a consequence point out that the gap approach implies that this consequence is a gap. Thus, the gap approach implies that one of its consequences is a gap. Hence, it implies that one of its consequences (i.e., (ρ)) is not true. Therefore, this theory of truth is self-refuting. Call this the *self-refutation reasoning*.[10]

Before continuing, I want to clarify what I take to be an important issue. If either of the above arguments is valid, then the 'not' in (ρ) must express exclusion negation. If we assume that it expresses choice negation, then the step from '(ρ) is a gap' to '(ρ) is not true' is invalid. If the 'not' in (ρ) is read as choice negation, then this inference would be equivalent to the inference from '(ρ) is a gap' to '(ρ) is false', and this inference is obviously invalid.

Let us distinguish between two versions of the liar sentence that figures in self-refutation problems:

(ρ1): (ρ1) is not_C true.

[9] See Burge (1979), Simmons (1991), and Gupta (2000).
[10] See McGee (1991).

(ρ2): (ρ2) is not$_E$ true.[11]

The gap approach certainly implies that (ρ1) is a gap. Neither version of the self-refutation problem (i.e., neither the strong liar reasoning nor the self-refutation reasoning) applies to this result. Given that (ρ1) is the natural reading of the standard liar sentence, the gap approach does a fine job of handling the standard liar.[12] The self-refutation problems arise when the gap approach confronts sentence (ρ2).

One should notice right away that when properly formulated, the strong liar reasoning is just the reasoning associated with the revenge paradox. Indeed, as I discuss below, (ρ2) is equivalent to (λ′) (on a strong Kleene reading of the disjunction in (λ′)). Thus, the strong liar reasoning poses no additional threat to the gap approach. On the other hand, the self-refutation reasoning might pose a new threat.

The gap theorist has at least two options for dealing with sentences like (ρ2). First, she can say that the gap theory is restricted so that it does not apply to (ρ2) (and sentences like it). Call this the *modest approach* to the self-refutation problem.[13] An adherent of the modest approach admits that her theory cannot deal with sentences like (ρ2). Of course, this sort of restriction renders any semantic theory for truth that is based on a gap theory of this sort essentially external for every truth language.[14]

[11] ‘not$_C$’ expresses choice negation and ‘not$_E$’ expresses exclusion negation.

[12] It is standard practice to assume that when extending sentential operators from a classical setting to a many-valued setting, the monotonic versions (e.g., ‘not$_C$’) are the most natural readings.

[13] The modest approach to the self-refutation problem and the modest approach to the revenge paradox are very similar. See Kripke (1975) for example.

[14] It might seem that adherents of the modest approach suffer from a new self-refutation problem that pertains to restricted theories. The modest gap theory is restricted so that no sentences containing exclusion negation operators are in its scope. Of course, (ρ2) contains an exclusion negation operator. Thus, it is outside the scope of the modest gap theory. Given that the modest gap theory is a theory of truth, it contains or has as a consequence a truth definition, which specifies which sentences are true, which ones are false, and which ones are gaps. Because (ρ2) is outside the scope of this truth definition, and the truth definition specifies which sentences are true, the truth definition implies that (ρ2) is not$_E$ true. That is, because the truth definition does not specify that (ρ2) is true, it

The second option for the gap theorist is to deny that (ρ2) is a genuine problem. That is, he claims that any language that contains a truth predicate has gaps and that exclusion negation is incoherent.[15] Thus, there is no sentence (ρ2) that causes a problem. Call this the *robust response* to the self-refutation problem. On the robust approach, the argument of the self-refutation problem is invalid. Thus, there is no reason to think that the gap approach implies that (ρ2) is not true. There is no way to express this claim according to the robust gap theorist.

The objector can respond to the robust gap theorist by constructing a new sentence:

(λ′): (λ′) is either false or a gap.

Of course, the gap theorist cannot deny the intelligibility of (λ′) because his own theory employs a gap predicate. As we saw in the discussion of the revenge problem, one can easily show that on the gap approach, (λ′) is true if and only if (λ′) is either false or a gap. The robust gap theorist can deal with this objection by claiming that (λ′) is a gap and that '(λ′) is true if and only if (λ′) is either false or a gap' is a gap as well. The robust gap theorist can justify these claims by appeal to certain readings of the logical connectives that occur in these sentences.[16]

The robust gap theorist who takes this path faces an additional objection. The robust gap theory implies that (λ′) is a gap. Thus, it implies that (λ′) is either false or a gap. Hence, (λ′) is a consequence of the robust gap theory. Then one can re-institute the self-refutation problem with (λ′); that is, the robust gap theory is self-refuting because it implies that one of its consequences

implies that (ρ2) is not_E true. Hence, the theory of truth on which the truth definition is based implies that (ρ2) is not_E true. Of course, '(ρ2) is not_E true' just is (ρ2), so the theory of truth has (ρ2) as a consequence. Hence, the theory of truth implies that one of its consequences is not_E true. Therefore, it is self-refuting. The problem with this objection is the claim that the theory of truth implies that (ρ2) is not_E true because (ρ2) is outside the scope of the theory. It is perfectly acceptable to say that a theory has no consequences for items that are outside its scope.

[15] See Parsons (1984), Tappenden (1999), and Maudlin (2004).

[16] See Blamey (2002) and Maudlin (2004).

is a gap. The robust gap theorist will have to agree with this assessment and respond by claiming that theories that are gaps are just as acceptable as theories that are true.

The robust gap theorist then faces two additional objections. First, the claim that it is acceptable to endorse a theory that is a gap causes problems with the notion of assertibility. Thus, it seems that the gap theorist has merely pushed off the problems posed by the liar from truth to assertibility. Hence, the robust gap theorist has done nothing more than move the bump in the rug into a dimly lit corner of the room in the hopes that no one will notice it there.[17] Instead of pushing this line of criticism, I want to consider a different problem with the robust gap approach. The robust gap theory implies that one of its consequences, (λ′), is a gap. However, it also implies that (λ′) is true because it implies that the truth rules are valid. Hence, the robust gap theory implies that (λ′) is both true and a gap. Hence, it implies that (λ′) is both true and not$_E$ true.

The robust gap theorist can respond to this charge with the claim that it does imply that (λ′) is a gap and it does imply that (λ′) is true, but that is not problematic because '(λ′) is true' is a gap as well. The robust gap theorist claims that, although somewhat counterintuitive, this result is not unacceptable; in fact, it is the only acceptable way of dealing with the liar.

Of course, the objector is not finished. The objector's next move is to either introduce into the object language logical connectives that allow him to formulate his objection properly or formulate the objection in a bivalent language (these are essentially the same move). The objection is then that the robust gap theory implies that (λ′) is true and that (λ′) is not true. Hence, the robust gap theory is inconsistent. The robust gap theorist is now backed into a corner and becomes aggressive. He claims that there is no such language in which to formulate the

[17] See Maudlin (2004).

objection. That is, it is illegitimate to introduce these logical connectives into the language because they are incoherent. Likewise, there are no bivalent languages that contain truth predicates. Thus, from the point of view of the robust gap theorist, the objection is either benign (if formulated in the gappy object language) or it is unintelligible (if formulated in what purports to be a bivalent language).

Instead of pursuing these objections to the robust gap theorist to this point, it seems to me that the opponent of the robust gap theory should take a stand with the robust gap theorist's first move. The robust gap theorist's first move is to deny that bivalent languages with truth predicates and languages with non-monotonic sentential operators are intelligible. Call this move the *unintelligibility response*. The objector should point out that it is unacceptable to defend one's theory by claiming that an entire topic of study (e.g., many valued logics with non-monotonic sentential operators) is incoherent or meaningless. (Notice that the same objection undermines the robust response to the revenge paradoxes discussed above.) It is obviously false that there are no bivalent languages with truth predicates or that the logical connectives used to formulate the objections are incoherent. Logicians have been studying such things for decades.[18] We can construct both bivalent languages with truth predicates and languages with non-monotonic sentential operators, and we can use them perfectly well. Moreover, provided that a bivalent language with a truth predicate cannot represent its syntax, it can be consistent; provided that a language with non-monotonic sentential operators does not contain a truth predicate, it can be consistent. Hence, the robust gap theorist's claim that none of these linguistic phenomena are intelligible is radically implausible.

[18] See Gupta (1982) for a discussion of consistent languages that contain truth predicates and Urquhart (2001) for an overview of many-valued logics.

Another problem with the unintelligibility response is that one cannot rationally deny that plenty of humans have been formulating and studying languages that contain non-monotonic sentential operators. To claim that these people have been dealing with meaningless expressions is absurd. The only semi-plausible view in the vicinity is that these expressions are meaningful but inconsistent. However, if the robust gap theorist is to be able to defend his claim that such things are inconsistent, then he owes us an account of what these people have been doing for the past fifty years. Any such account will have to deal with the paradoxes that can result from such devices (i.e., the revenge paradoxes). One might as well just use such an account for the liar paradox and avoid the radically implausible claim that all non-monotonic sentential operators and all bivalent languages with truth predicates are incoherent. My conclusion is that a robust gap theory is unacceptable—the self-refutation problem is a genuine concern for theories of truth that respect the truth rules and have paradoxical sentences within their scope.

So far, I have discussed the revenge problems and the self-refutation problems; I have also presented both a modest and a robust response to each. Now I argue that any theory of truth that respects the truth rules faces either a revenge paradox or a self-refutation problem unless it is restricted or incorporates the unintelligibility response. The heart of the argument is showing that the revenge paradoxes and the self-refutation problems reinforce one another—attempts to avoid one bring on the other; one can avoid both of them only by restricting one's theory or endorsing an unintelligibility response.

Consider the theory of truth T that accepts the truth rules and classical logic. The standard liar sentence (λ) poses a problem for T. Either T implies that (λ) is both true and false or T is restricted so that sentences like (λ) are not in its scope. Thus, T is either inconsistent or restricted.

Now consider a theory of truth T′ that respects the truth rules and implies that (λ) is a gap (i.e., (λ) is not a member of the extension of ‘true’ and (λ) is not a member of the anti-extension of ‘true’). Given that T′ treats truth as partially defined, it must abandon classical logic for some many-valued logic. An argument of the same form that shows (λ) causes a problem for T shows that (λ′) causes a problem for T′. Below are both arguments:

(a)	(λ) is true. (Assumption)	(a′)	(λ′) is true.(Assumption)
(b)	‘(λ) is false’ is true. (Sub)	(b′)	‘(λ′) is false or a gap’ is true. (Sub)
(c)	(λ) is false. (Desc)	(c′)	(λ′) is false or a gap. (Desc)
(d)	(λ) is false. (Assumption)	(d′)	(λ′) is false or a gap. (Assumption)
(e)	‘(λ) is false’ is true. (Asc)	(e′)	‘(λ′) is false or a gap’ is true (Asc)
(f)	(λ) is true. (Sub)	(f′)	(λ′) is true. (Sub)

Thus, if T and T′ accept the truth rules and T implies that (λ) is true iff (λ) is false, then T′ implies that (λ′) is true iff (λ′) is false or a gap.

There are seven possibilities for the way T′ handles (λ′):

(1) T′ implies that ‘(λ′) is true iff (λ′) is false or a gap’ is a contradiction. Hence, (λ′) constitutes a revenge paradox for T′, and T′ is inconsistent.

(2) T′ implies that the gap predicate in (λ′) is partially defined and (λ′) is a ‘gap’ gap. Hence, T′ incorporates the robust response to the revenge paradox. However, one can construct a new sentence that is just like (λ′) except that it contains a completely defined gap predicate. Then T′ will imply that this new sentence is both true and either false or a gap.

(3) T′ incorporates the robust response to the revenge paradox just as in the second option, but it incorporates an unintelligibility response by claiming that completely defined gap predicates and any other linguistic devices that can be used to construct genuine revenge paradoxes are incoherent. Thus, T′ is false because it implies that certain coherent linguistic devices are incoherent.

(4) T′ implies that ‘(λ′) is true iff (λ′) is either false or a gap’ is a gap. Hence, T′ implies that one of its consequences is a gap. Therefore, (λ′) constitutes a self-refutation problem for T′, and T′ is unacceptable.

(5) T′ implies that ‘(λ′) is true iff (λ′) is either false or a gap’ is a gap. Hence, T′ implies that one of its consequences is a gap. However, T′ implies that gappy theories are acceptable. Thus, T′ incorporates the robust response to the self-refutation problem.

However, one can argue that because T′ respects the truth rules, it implies that its consequence is both true and not$_E$ true. Therefore, T′ is false.

(6) T′ incorporates the robust response to the self-refutation problem just as in the fifth option, but it incorporates an unintelligibility response by claiming that non-monotonic sentential operators and any other linguistic devices that can be used to formulate the self-refutation objection are incoherent. Therefore, T′ is false because it implies that certain coherent linguistic devices are incoherent.

(7) T′ is restricted so that sentences like (λ′) in which gap predicates occur are not in its scope.

Of these seven options, only the seventh results in an acceptable theory. Of course, it also results in a theory of truth that is restricted from applying to languages or sentences with gap predicates. Because it requires a gap predicate for its formulation, it is essentially external for every language. Therefore, the only acceptable theory that respects the truth rules is essentially external for every language.

I want make several comments on the other options and on the argument as a whole before moving on to the next section. One important thing to notice about the above discussions of the revenge problem and the self-refutation problem is that they concern the same sentence: (λ′). For some gap theories, (λ′) figures in a revenge paradox, while for others, it figures in a self-refutation problem. Another key observation is that no matter what truth-status T′ assigns to (λ′), T′ has (λ′) as a consequence. If T′ implies that (λ′) is true, then T′ has (λ′) as a consequence. If T′ implies that (λ′) is false, then T′ implies that (λ′) is either false or a gap; hence, T′ has (λ′) as a consequence. If T′ implies that (λ′) is a gap, then T′ implies that (λ′) is either false or a gap; hence T′ has (λ′) as a consequence. If T′ implies that (λ′) is a 'gap' gap, then there is some other sentence that is just like (λ′) but contains a completely defined gap predicate; if we call that other sentence (λ′), then T′ has (λ′) as a consequence.

The fact that the items that cause trouble for a theory of truth are sentences and, moreover, that they are consequences of the theory itself makes the liar paradox especially difficult.[19] It is this fact that allows the revenge paradoxes and the self-refutation problems to "team up" against theories of truth that accept the truth rules.

In short, the problem is this: any theory of truth that respects the truth rules will have to classify (λ'), and it has (λ') as a consequence. If it classifies (λ') as true or false, then it is inconsistent because it has a contradiction as a consequence (i.e., (λ') poses a revenge paradox). If it classifies (λ') as a gap, then it is self-refuting because it classifies one of its consequences as a gap (i.e., (λ') poses a self-refutation problem). No wonder some theorists are driven to claim that the linguistic resources required to formulate a revenge paradox or a self-refutation objection are unintelligible; it seems to be the only way to avoid both the revenge paradox and the self-refutation problem other than restricting one's theory. The theorist who endorses the truth rules is dining at a bullet buffet.

3.3 Truth, Inconsistency, and Internalizability

In the previous section, I presented an explanation for the dearth of internalizable semantic theories for truth. I claim that the solution to this problem is a theory of truth on which truth is an inconsistent concept. Such a theory cannot have the form of most theories of truth, which are just the principles that govern the concept of truth. Such theories are obviously inconsistent. What is needed is a consistent theory of truth that treats truth as an inconsistent concept. That is the goal for the rest of the dissertation. In this section, I present four arguments for treating truth

[19] Other well-known paradoxes (e.g., Russell's paradox, Grelling's paradox, Berry's paradox) do not have this feature. Thus, they do not pose the same difficulty we see with the liar paradox.

as an inconsistent concept. My goal is not to *prove* that truth is an inconsistent concept. Rather, I aim to give good reasons for pursuing theories of truth that treat truth as an inconsistent concept.

The first argument is based on the discussion in the previous section. The argument there shows that if a theory of truth accepts the truth rules, then that theory of truth is either inconsistent, self-refuting, restricted, or unacceptable (because it includes an unintelligibility response). That is, any unrestricted theory of truth that respects our intuitions about truth will be unacceptable. One lesson to learn from this is that the principles we take to govern our concept of truth really are inconsistent; hence, if our concept of truth behaves as we think it does, then our concept of truth is inconsistent. That is, the principles that are constitutive of this concept are inconsistent. Another lesson is that an acceptable theory of truth cannot take the form of a collection of principles that govern truth. If truth is an inconsistent concept, then any such theory will be inconsistent.

One might try to avoid these conclusions by claiming that our concept of truth is not as we take it to be. There are two broad categories of theories that deny the truth rules. Those in the first category posit some hidden semantic feature of truth in order to avoid the liar paradox and claim that the truth predicate obeys rules that are very similar to the truth rules but are specific to concepts with this hidden semantic feature. The second type of theory flatly denies that the truth rules are valid. I consider them in order.

There are several theories of the first type. Ambiguity theories claim that truth predicates in natural languages are ambiguous. Usually, these theories imply that natural language truth predicates are *radically* ambiguous. By that, I mean that they imply that instead of common instances of ambiguity (e.g., ‘bank’), a natural language truth predicate can express a transfinite number of distinct concepts. Usually, such theories posit restricted versions of the truth rules for each of the distinct meanings of the truth predicate. I discuss these theories in detail and pose several objections to them in Appendix A.

Contextual theories of truth also posit a hidden semantic feature in our natural language truth predicate. They claim that sentences containing natural language truth predicates display the same sort of context dependence as those containing indexicals (e.g., ‘here’), attributive adjectives (e.g., ‘tall’), or quantifiers (e.g., ‘everything’). They also formulate rules that are similar to the truth rules, but are specific to context dependent expressions. I discuss these theories and pose several criticisms of them in Appendix B.

The revision theory of truth is another theory that posits a hidden semantic feature. I have already discussed this theory several times. It implies that truth is a circular concept that is governed by a rule of revision instead of the usual rule of application that governs most concepts. Again, the revision theory implies that truth obeys principles that are similar to the truth rules, but are specific to circular concepts.[20]

All of the theories of truth that posit hidden semantic features in our natural language truth predicates face at least three major problems. First, they imply that our truth predicates have certain features that have gone unnoticed despite centuries of use and study. These theories posit these features purely as a way of avoiding the liar paradox and its kin. Thus, there is no independent reason to think that our truth predicates have these features. Second, these theories face revenge paradoxes of their own. For example, a contextualist theory has just as much trouble with a sentence like (κ) (where (κ) is ‘(κ) is false in all contexts’) as the gap theory has

[20] Another example is Skyrms’ intensional theory, which denies the substitution truth rule; see Skyrms (1970a, 1970b, 1984).

with (λ').[21] Thus, there is no real payoff for positing hidden semantic features of natural language truth predicates. Third, these theories miss the real point of an approach to the liar paradox. The problem is not: what should we say about truth given that it generates these terrible paradoxes? The problem is: what should we say about these terrible paradoxes? Positing some hidden semantic feature of natural language truth predicates might save our natural language truth predicates, but so long as one can just introduce a linguistic expression that behaves like a truth predicate without this hidden feature, it does not offer a solution to the liar paradox. It simply reclassifies the liar paradox as a problem facing some concept other than truth. Saying that the liar paradox is not a problem for natural language truth predicates because they do not behave in the way that generates the paradox does nothing to answer the question of what we should do about the paradoxes that result from linguistic expressions that *do* behave in the way that generates the paradox.[22] These theories leave us with implausible theories of truth and no approach to the liar paradox.

Another way of avoiding the conclusion that an inconsistency theory of truth is the only acceptable approach to truth is to deny flat out one or more of the truth rules.[23] Of course, the most obvious reason to accept the truth rules is that they seem to be constitutive of the concept of truth. One problem facing theories of truth of this sort is that most people will not take them to be theories of *truth*. A concept that does not obey the truth rules is a concept that is distinct from our concept of truth.

Even if I could be persuaded that our concept of truth does not obey the truth rules, I would still have a problem with these theories *qua* approaches to the liar paradox. They face the same problem that confronts those who posit hidden semantic features of natural language truth

[21] In Appendix C, I argue that the revision theory faces revenge paradoxes as well.
[22] Yablo (1993b) makes this point as well.
[23] See Feferman (1982) and Reinhardt (1986) for examples. See Friedman and Sheard (1987) for discussion.

predicates: even if their claims about truth are plausible, at most they show that the liar paradox is not a paradox that pertains to truth. These theories do not constitute approaches to the liar paradox because one can simply introduce an expression for which the truth rules *are* constitutive. Thus, denying that the truth rules hold is analogous to renaming the liar paradox to indicate that, although it still causes horrible problems, those horrible problems do not concern the concept of truth. Given that truth seems to obey these rules and that we need an approach to concepts that obey these rules anyway, it is better to propose an account of such concepts and apply that account to truth. Assuming that the argument of the previous section is correct, it is impossible to provide a theory of truth that: (i) respects the truth rules, (ii) is unrestricted, (iii) treats truth as a consistent concept, and (iv) is not obviously false (e.g., inconsistent, self-refuting, or incorporates an unintelligibility response). The way out of this predicament is to treat truth as an inconsistent concept.

The second argument is that if one decides to treat truth as an inconsistent concept, then one has available a satisfying explanation of the current situation in truth studies. That is, one can explain why other theories of truth face either self-refutation problems or revenge paradoxes and so are unable to serve as the basis for internalizable semantic theories for truth. No other theory of truth has managed to do this.[24]

The explanation for why theories of truth that imply truth is a consistent concept face revenge paradoxes or self-refutation problems is straightforward. Our concept of truth is inconsistent in the sense that its constitutive principles, the truth rules, are incompatible. That is, there are objects that these rules classify as both true and not true. We can call the set of such objects the *overdetermination set* for truth. All the paradoxical sentences considered so far are members of the overdetermination set for truth. Any naïve theory of truth that includes these

[24] See Glanzberg (2005) for the only other explanation of which I am aware.

principles and accepts classical logic is inconsistent and can be rendered consistent only by restricting it. If a theory of truth implies that some of the sentences in the overdetermined set for truth are gaps, then its fate depends on which of these sentences it classifies as gaps. Recall that many of the members of the overdetermination set for truth are truth status attributions, and no matter what truth status one assigns them, they are consequences of the assignment. No matter whether one's theory of truth classifies these paradoxical sentences as true, false, or gappy, these paradoxical sentences are consequences of one's theory. Thus, if a theory of truth implies that all the sentences in the overdetermined set for truth are gaps, then the theory implies that some of its consequences are gaps. On the other hand, if a theory of truth does not classify some of these sentences as gaps, then the truth rules imply that they are both true and not true. On the first option, the theory is self-refuting, while on the second, it is rendered inconsistent by a revenge paradox.

In section two, I argued that theories of truth that accept the truth rules are not internalizable for natural languages because they face either revenge paradoxes or self-refutation problems. If we admit that truth is an inconsistent concept, then we can explain why such theories inevitably face either revenge paradoxes or self-refutation problems. Therefore, by accepting that truth is an inconsistent concept, we arrive at a deeper explanation for why theories of truth that accept the truth rules fail to be internalizable for natural languages.

The third reason for adopting an inconsistency theory of truth is that Vann McGee has shown that a theory of truth that entails the obvious principles governing the interaction of the truth predicate with the logical expressions of classical first order logic is ω-inconsistent. More precisely, McGee's theorem states that a set, S, of sentences that meets the following conditions is ω-inconsistent:

(a) S contains the axioms of arithmetic.
(b) S is closed under first-order consequence.
(c) S contains ⟨⟨p⟩ is true⟩ whenever it contains ⟨p⟩.
(d) S contains all instances of: ⟨⟨p implies q⟩ is true⟩ implies ⟨⟨p is true⟩ implies ⟨q is true⟩⟩.
(e) S contains all instances of: ⟨⟨~ p⟩ is true⟩ implies ⟨~ ⟨p⟩ is true⟩.
(f) S contains all instances of: ⟨For all x, ⟨p(x)⟩ is true⟩ implies ⟨⟨for all x, p(x)⟩ is true⟩.

A set is ω-inconsistent (roughly) if and only if there is some formula ⟨p(x)⟩ such that ⟨it is not the case that for all x, p(x)⟩ is a consequence of the set of sentences, but for each n, ⟨p(n)⟩ is a consequence of it. Thus, a set is ω-inconsistent if it implies all the instances of a generalization, and implies the negation of the generalization. Granted, McGee's theorem does not show that a theory of truth that satisfies these properties is inconsistent (indeed, he proves that his result cannot be strengthened in this way), but ω-inconsistency is bad enough. For example, one can show that an ω-inconsistent theory cannot give arithmetical expressions their normal meanings.[25] My point is that the conditions in McGee's theorem are natural ones to want for a theory of truth. He has shown that any naïve theory of truth that satisfies them is unacceptable (because it is ω-inconsistent). Thus, we have good reason to think that any theory of truth on which truth is a consistent concept will be implausible because it either denies one of these eminently plausible principles or it is ω-inconsistent. That gives us another reason to prefer an inconsistency theory of truth.

The fourth and final reason for adopting an inconsistency theory of truth is that by doing so, one can avoid the self-refutation problems and revenge problems that plague the other theories of truth. Thus, an inconsistency theory of truth does not need to be restricted to avoid these problems. Hence, it can serve as the basis for a semantic theory for truth that is internalizable for every language. If that is a legitimate goal, and I have argued at length in

[25] See McGee (1985).

Chapters One and Two that it is, then the prospect of an internalizable semantic theory for truth is one of the best reasons for adopting an inconsistency theory of truth. In other words, if (i) only theories of truth that support internalizable semantic theories for truth really explain truth, (ii) an inconsistency theory of truth can support an internalizable semantic theory for truth, and (iii) we have good reason to believe that a consistency theory of truth cannot support an internalizable semantic theory for truth, then we have good reason to believe that only an inconsistency theory of truth really explains truth. Of course, I have not yet shown that by admitting that truth is an inconsistent concept one can construct a theory of truth that serves as the basis for an internalizable semantic theory for truth. However, the theory I develop in Chapters Four, Five, Six, and Seven justifies this claim. It is a consistent theory of truth that faces no self-refutation problems and no revenge paradoxes; hence, it does not require restrictions, and it can serve as the basis for an internalizable semantic theory for truth (which I also provide).

3.4 Conclusion

In this chapter, I have argued for treating truth as an inconsistent concept. In section 3.2, I presented detailed accounts of the self-refutation problem and the revenge paradoxes. I argued that any theory of truth that treats truth as univocal and invariant and respects the truth rules is self-refuting, inconsistent, restricted, or radically implausible. In section 3.3, I presented four arguments for treating truth as an inconsistent concept. Thus, this chapter serves as motivation for the approach I develop in the rest of the dissertation.

4.0 A Theory of Inconsistent Concepts

4.1 Introduction

This chapter is the first of three in which I present a theory of inconsistent concepts. In this chapter, I propose a definition of 'inconsistent concept' and present several distinctions and examples. I also discuss three policies for dealing with a concept one has discovered to be inconsistent, and I argue that inconsistent concepts should be replaced. Finally, I discuss the link between conceptual inconsistency and conceptual confusion, which is the focus of Chapters Five and Six.

4.2 Inconsistent Concepts

In this section, I provide a definition of 'inconsistent concept' and discuss several distinctions and examples; I also present my account of inconsistent concepts. Before discussing inconsistent concepts I want to make a few remarks about concepts. I prefer to accommodate a range of views on the nature of concepts. I do invoke the expression relation that holds between words and concepts; however, I provide no explicit account of it. I also speak of applying a concept. Roughly, I think of concept application on the lines of belief formation or assertion. For example, I apply the concept **scab** to some object α if I am prepared to assert 'α is a scab' or I believe that α is a scab.[1]

An important part of my account of concepts is that there are rules that govern the employment of a concept. A person who possesses a certain concept and is committed to employing it is committed to following the rules for the employment of that concept. One such rule is that the concept **scab** should be applied to scabs and it should not be applied to things that are not scabs.[2] One can think of these rules as *constitutive* in the sense that if a person utters a word that expresses the concept **scab**, then that person is committed to following the rules for the employment of **scab**. By 'committed to following the rules' I do not mean that the person actually acts in accordance with these rules or that he explicitly endorses them; rather, I mean that the person *ought* to follow them—he is obligated to follow them—whether he explicitly endorses them or not.[3]

An *inconsistent concept* is one whose constitutive rules are incompatible in the sense that they dictate that the concept both applies and does not apply to some entities. The rules for the employment of an inconsistent concept impose conflicting commitments on the employers of that concept. Thus, the employer of an inconsistent concept cannot follow the rules for the application of that concept in all circumstances.[4] Consider an example:

(1a) 'rable' applies to x if x is a table.

(1b) 'rable' disapplies to x if x is red.[5]

[1] I use bold type as a convention for the names of concepts.

[2] I am not committed to explaining concepts in terms of such rules; see Davidson (1982) for a criticism of those who favor such an explanatory strategy.

[3] My commitment to constitutive rules for concepts places me in the tradition of meaning-constitutive accounts of concepts. However, not all the members of this tradition agree on rules as the relevant constitutive element. Some constitutive accounts choose the possession of propositional attitudes, the truth of theories, the validity of implications, etc. See Peacocke (1992) for an example.

[4] See Chihara (1979, 1982) and Yablo (1993a) for similar views on inconsistent concepts.

[5] I say that a concept *applies* to the members of its extension and *disapplies* to the members of its antiextension.

Rable is an inconsistent concept. Someone who possesses **rable** might run into difficulty employing it because it both applies and disapplies to red tables. When confronted with a red table, an employer of **rable** will be unable to satisfy the demands it places on her. Of course, someone could employ **rable** without trouble as long as she avoids red tables. Notice that **rable** is undefined for things that are neither red nor tables.

It is essential to distinguish between inconsistent concepts and unsatisfiable concepts. An *unsatisfiable concept* is one that is consistent but which does not apply to anything. An unsatisfiable concept places incompatible demands on the objects to which it applies, while an inconsistent concept places incompatible demands its employers. For example,

(2) x is a *squircle* if and only if x is a square and x is a circle.

Squircle is an unsatisfiable concept, but it is not inconsistent. Someone who possesses **squircle** has no problem employing it. It should be disapplied to everything.[6]

I want to make several points about inconsistent concepts. First, attempting to place the definition of an inconsistent concept in the standard form results in a consistent concept that is either conjunctive or disjunctive. Notice the difference in definitions (1) and (2). (2) prescribes both the application conditions and the disapplication conditions for **squircle** at once, while (1) has two separate clauses for **rable**. When considering a definition like (2), it is common to assume that if something is not both a square and a circle, then it is not a squircle. This assumption fits well with consistent concepts because their application conditions and disapplication conditions are disjoint. The application conditions and disapplication conditions for inconsistent concepts overlap. That makes it impossible to introduce them with definitions that are in the form of (2). Consider another definition:

[6] I mention the distinction between inconsistent and unsatisfiable concepts because it is a common mistake to assume that inconsistent concepts are merely unsatisfiable. See Stenius (1972), Chihara (1979), and Yablo (1993b) for discussions of the distinction and the mistake.

(3) x is a *non-red-table* if and only if x is a table and x is not red.

There is a big difference between **non-red-table** and **rable**. **Non-red-table** is consistent and applies to things that are both tables and not red; it disapplies to everything else.

Second, inconsistent concepts characteristically give rise to paradoxes. For example, assume for reductio that some red tables exist. Let *R* be the name of a red table. R is a table; hence, R is a rable. R is red; hence, it is not the case that R is a rable. Thus, R is a rable and it is not case that R is a rable. Contradiction. Therefore, no red tables exist. It is obvious that something has gone wrong in this argument, but what? I take it as a condition on any account of inconsistent concepts that it must explain the fallacy in the above argument. It should not be surprising that arguments like this one feature prominently in criticisms of theories that posit inconsistent concepts. I address it in Chapter Six.

The next point is that there is an affinity between inconsistent concepts and partial concepts. A *partial concept* is one that has a limited range of applicability. Some concepts are partial by definition. Here is Scott Soames' example of a partial concept:

(4a) 'smidget' applies to x if x is greater than four feet tall;

(4b) 'smidget' disapplies to x if x is less than two feet tall.[7]

Smidget is a partial concept because it is undefined for entities that are between two and four feet tall. I want to introduce several terms that are helpful in discussing partial concepts and inconsistent concepts. When discussing any partial concept, I assume that there is a set of all the objects that exist; I call it the *domain*. This assumption brings with it several obvious and difficult set theoretic problems that I will not go into; they do not matter for my purposes. I say that the *range of applicability* of a concept is the subset of the domain to which it either applies or disapplies. The *range of inapplicability* is usually the complement of the range of

[7] Soames (1999). See Glanzberg (2003) for criticism.

applicability. I say that a concept is *inapplicable to an object* if that object falls within its range of inapplicability. **Smidget**'s range of applicability is the set of objects that are either greater than four feet tall or less than two feet tall. **Rable**'s range of applicability is the set of objects that are either tables or non-red things. For the purposes of distinguishing between inconsistent and consistent concepts, I draw a distinction between the objects a concept applies to and those it is true of. Both inconsistent and consistent concepts can apply to objects, but only consistent concepts can be true of objects. This usage coincides with my conclusion in Chapter Five that sentences involving inconsistent concepts are truth-value gaps. It would sound odd to say that Φ is true of α, but $\ulcorner\Phi\alpha\urcorner$ is not true. Likewise, only consistent concepts have extensions or antiextensions. I call the set of things to which a concept applies its *application set* and the set of things to which a concept disapplies is its *disapplication set*. The application sets of consistent concepts are their extensions and the disapplication sets of consistent concepts are their anti-extensions. A concept's *overdetermined set* is the intersection of its application set and its disapplication set. One must be especially careful dealing with negation and partial concepts. 'α is not a smidget' is ambiguous because it can either mean that smidget disapplies to α or that smidget is inapplicable to α. The former reads 'not' as choice negation and the latter reads it as exclusion negation (I discuss these issues at length in Chapter Seven).

There is an important distinction between *completely defined partial concepts* and *partially defined partial concepts*.[8] The former have explicitly specified ranges of applicability and inapplicability. The latter have explicitly specified ranges of applicability, but nothing is said about their range of inapplicability. **Smidget** is a partially defined partial predicate. The following definition of 'smidget' makes it completely defined partial predicate:

[8] Soames (1999) makes use of this distinction in his theory of truth. See Gupta (2002) for this terminology, which Soames endorses in Soames (2002b).

(5a) ‘smidget’ applies to x if x is greater than four feet tall;

(5b) ‘smidget’ disapplies to x if x is less than two feet tall;

(5c) ‘smidget’ is inapplicable to x if x is between two and four feet tall (inclusive) or x does not have a height.

If a concept’s application set and disapplication set are neither disjoint nor jointly exhaustive, then it will be both partial and inconsistent.

Up to this point I have discussed only inconsistent concepts whose application sets and disapplication sets are not disjoint. However, if a concept’s range of applicability and its range of inapplicability are not disjoint, then it is inconsistent as well. For example:

(6a) ‘mammamonkey’ applies to x if x is a mammal.

(6b) ‘mammamonkey’ disapplies to x if x is an animal and x is not a mammal.

(6c) ‘mammamonkey’ is inapplicable to x if x is either a monkey or x is not an animal.

Although the application set and disapplication set for **mammamonkey** are disjoint, it is an inconsistent concept because its range of applicability and range of inapplicability overlap. A concept can exhibit both types of inconsistency as well. I mark this distinction by saying that an *application-inconsistent* concept is one whose application set and disapplication set are not disjoint; a *range-inconsistent* concept is one whose range of applicability and range of inapplicability are not disjoint. I focus primarily on application-inconsistent concepts in the remainder of this chapter, but most of my comments and results hold for range-inconsistent ones as well.

Before discussing policies for handling conceptual inconsistency, I present several other types of inconsistent concepts. I have already introduced **rable**, which is inconsistent and causes problems for anyone who decides to employ it in the vicinity of red tables. However, even if an employer of **rable** never encounters a red table, the concept still poses a problem for her because inconsistent concepts pose a normative problem for their employers. Someone who chooses to employ **rable** *should* apply it to tables and *should* disapply it to red things. These are conceptual norms to which the employer has decided to bind herself. Of course, a concept possessor can decide to employ a certain concept without knowing that it is inconsistent. All the types of inconsistent concepts I discuss cause normative problems for their employers; however, they differ on the availability of the items that cause problems for someone who has decided to employ the inconsistent concept. Red tables are plentiful; hence, an employer of **rable** will run into trouble in pretty common circumstances.

It is possible to define an inconsistent concept that can be employed without difficulty in any physically possible situation. Consider the following definition:

(7a) 'uranicube' applies to x if x is a cube whose volume is at least one cubic mile.

(7b) 'uranicube' disapplies to x if x is composed entirely of uranium.

As I have defined it, **uranicube** is both partial and inconsistent. Its range of applicability is the union of the set of cubes whose volumes are greater than one cubic mile and the set of things composed entirely of uranium. I assume that, according to the laws of nature, it is physically impossible for a cube of pure uranium whose volume is at least one cubic mile to exist (this is a stock example from philosophy of science discussions). Thus, an employer of **uranicube** will not run into any difficulty while applying it to objects of the actual world. Although **uranicube** is an inconsistent concept and an employer of it is in normative difficulty, he will never have to decide whether to apply it or disapply it to an object in its overdetermination set.[9]

Even less threatening inconsistent concepts are possible as well. Consider:

(8a) 'cirquare' applies to x if x is a square.

[9] Depending on one's views on counterfactuals and laws of nature, an employer of **uranicube** might run into trouble by using it in certain subjunctive conditionals or by formulating natural laws with it.

(8b) ‘cirquare’ disapplies to x if x is a circle.

Cirquare is an inconsistent concept, but since squircles are conceptually impossible, there is no actual or possible threat to an employer of it. In fact, **cirquare** is so benign that one might deny that it is inconsistent at all.

Up to this point I have discussed mostly inconsistent concepts that are inconsistent *by definition*. However, one can construct an example of a concept that is inconsistent by virtue of the environment in which it is used. I call the former *intrinsically inconsistent* and the latter *empirically inconsistent*. The following is an example of an empirically inconsistent concept based on an example of Gupta’s.[10]

Consider a community of people who speak a language that is similar to English except that in their language, the rules for using the expression ‘x is up above y’ (where ‘x’ and ‘y’ are replaced by singular terms) are different. I call the members of this community *Higherians*. Two equally important features of the Higherians’ ‘up above’ talk are that they can perceptually distinguish situations in which one object is up above another (these situations are similar to the ones in which an English speaker would say that one object is up above another), and that they can determine when the ray connecting two objects is parallel to a particular ray that is designated as “Standard Up” (where Standard Up is orthogonal to a tangent plane for the surface of the object on which the Higherians live). An assertion of ‘A is up above B’ is warranted if and only if either A and B are constituents of one of the perceptually distinguishable situations (call this the *perceptual criterion*), or the ray connecting A and B is parallel to Standard Up and A is farther from the surface than B (call this the *conceptual criterion*). An assertion of ‘A is not up above B’ is warranted if and only if either A and B are not in the proper perceptually

[10] Gupta (1999).

distinguishable relation to one another, or it is not the case that both the ray connecting A and B is parallel to Standard Up and A is further from the surface than B.

Assume that ‘up above’ is defined only for perceivable objects and only for objects within the national borders of the Higherians’ country. When a Higherian can perceive two objects at the same time then that person can perceive whether they are in the right perceptually distinguishable relation to one another. In addition, every Higherian can determine the ray that connects any two perceivable objects and can determine whether any two rays are parallel. Thus, if a Higherian can perceive object A and he can perceive object B (not necessarily simultaneously), then he can determine whether the ray that connects them is parallel to Standard Up. Assume that the Higherians do not know that their concept is inconsistent because when they can perceive two objects at the same time, they employ the perceptual criterion and when they cannot, they employ the conceptual criterion. Assume also that whether one object is up above another does not depend on any of the Higherians taking them to be in this relation and that the notion of warrant is not relative to anyone’s epistemic situation. Finally, assume that there is no difference between the Higherians’ idiolects and their common language, that there is no conversational implicature associated with statements containing ‘up above’, and that the conventions governing ‘up above’ are common knowledge (i.e., there is no division of linguistic labor for this expression).

If the Higherians live on the surface of a spherical planet, and their nation consists of more than just a single point, then ‘up above’ is inconsistent. If A and B are two objects that are located some distance from where Standard Up intersects the surface of their sphere and are in the right perceptually distinguishable relation then both ‘A is up above B’ and ‘A is not up above B’ will count as warranted because they are in the right perceptually distinguishable relation, but the ray connecting them is not parallel to Standard Up. However, if the Higherians' country is confined to one flat surface of a rectangular solid, then 'up above' is consistent because it is defined only within their national borders. Hence, **up above** is an empirically inconsistent concept in the case where the Higherians live on the surface of a sphere.

It might seem impossible for a concept to be inconsistent without the employers of that concept knowing that it is inconsistent, but empirically inconsistent concepts should dispel this impression. The rules for the employment of a concept often incorporate features of the environment in which it is used; if the employers of a concept are ignorant or mistaken about some features of their environment, then the concept in question can be inconsistent without their knowledge. No amount of "reflection on their concepts" will inform them that their concept is inconsistent; they have to go out into the world and learn empirical facts to discover the conceptual inconsistency. Consider the history of human inquiry—we (humans) discover false empirical beliefs alarmingly often. Given the degree of our ignorance and error, there is a good chance that many, perhaps most, of our concepts are empirically inconsistent. That sobering thought should lend urgency to the task of constructing an adequate theory of inconsistent concepts and a descriptively complete and descriptively correct semantic theory for inconsistent concepts.[11]

4.3 Policies for Handling Inconsistent Concepts

What should a person do if she discovers that she employs an inconsistent concept? What should a person do if he discovers that someone else employs an inconsistent concept? I address the

[11] One can construct a concept that is inconsistent by virtue of the natural laws of the world in which it is used (e.g., the pre-relativistic concept of **simultaneity**). I suggest the term 'nomically inconsistent' for such concepts.

second question in Chapters Five and Six. Here I want to discuss three potential answers to the first one. To have a concrete example, assume that Troy is a person who has discovered that one of his dearly beloved concepts, concept X, is inconsistent. I do not address the difficult issue of how one discovers such a thing.

4.3.1 The Reinterpretation Policy

Suspicions of conceptual inconsistency are invariably accompanied by efforts to reinterpret the conceptual employment in question. The reinterpretation policy makes this reaction the official strategy for dealing with inconsistent concepts. In my example, one thing Troy could do is reinterpret his past actions and beliefs so that either he never employed X or X is not inconsistent. The first option leaves X alone and posits a consistent concept, Y, as the one Troy was using all along. The second option reinterprets X. The sort of reinterpretation I have in mind here is similar to the maneuvers found in Kripke's rule-following argument[12], Quine's argument for indeterminacy of translation[13], and Goodman's new riddle of induction[14]. I do not doubt that such a reinterpretation is possible, but I do question the legitimacy of the reinterpretation policy as a way of dealing with the discovery of an inconsistent concept. It seems to me that it would not be hard to construct situations of inconsistent concept employment that would force very strange and uncharitable reinterpretations. It is far better to have an account ready to hand that can be used in the event of such a discovery. I want to emphasize that I have no argument to show that the reinterpretation strategy is impossible or that people do not

[12] Kripke (1982) contains an example where someone reinterprets **plus** as **quus**.

[13] Quine (1960) contains an example where someone reinterprets **rabbit** as **undetached-rabbit-part** or **rabbit stage**.

[14] Goodman (1955) contains an example where someone reinterprets **green** as **grue**.

use it. On the contrary, it is the most common response. My qualm is with having it as a general strategy for dealing with inconsistent concepts. I discuss it more in Section 6.7 of Chapter Six.

4.3.2 The Containment Policy

According to the containment policy, we should identify the overdetermined items for concept X and treat them in a way so as to render them benign. That means we should determine which objects are in X's overdetermination set and avoid applying or disapplying X to them. We should refrain from asserting sentences associated with these employments of X and avoid having propositional attitudes associated with them as well. (For example, if R is a red table, then we should assert neither 'R is a rable' nor 'R is not a rable' and we should believe neither that R is a rable nor that R is not a rable. A number of prominent philosophers have advocated one form or another of the containment policy for dealing with the aletheic paradoxes (e.g., the liar), including Katsoff, Popper, van Bentham, Chihara, and Yablo.[15] It is also a common view among non-philosophers who are presented with paradoxes that arise in connection with inconsistent concepts.

I have several reservations about the containment policy. First, it can turn out to be difficult or impossible to avoid either uttering paradoxical sentences or entertaining attitudes toward the propositions they express (if such propositions are possible). In the case of truth, the knowledge required to determine whether a particular sentence is paradoxical goes far beyond that which any normal speaker has in everyday situations, and in some cases it is beyond anyone's knowledge. Of course, sometimes we can figure out whether a sentence is paradoxical, but in general, determining whether any given sentence is paradoxical is incredibly difficult

[15] See Katsoff (1953), Popper (1954), van Bentham (1978), Chihara (1979, 1984), Yablo (1985, 1989).

because it can depend on the semantic properties of sentences to which we no longer have access. Likewise, in many cases it is far too demanding to restrict people from applying or disapplying an inconsistent concept to paradoxical items (see Appendix B for a discussion of this issue). These problems take on a prominent role in the case of truth in Chapter Seven.

My biggest concern about the containment policy is that does not get to the root of the problem. The problem is that an inconsistent concept is defective. There is a sense of 'ought' in which defective concepts ought to be replaced or retired. It is the same sense as the 'ought' in 'one ought to avoid moral dilemmas'. Because the employer of an inconsistent concept undertakes incompatible commitments (i.e., undertakes commitments to follow the rules for employing the concept in question), one ought to refrain from employing an inconsistent concept. However, a proponent of the containment policy tries to deal with the problems that result from the continued employment of the inconsistent concept. In an important sense, the containment policy treats the symptoms instead of the disease.

I think that there is a limited place for the containment policy in an effective paradoxicality response program. The containment policy is an important first step in the eventual replacement of an inconsistent concept. While we (humans) are deciding what changes to make to our conceptual repertoire, the containment policy is the best one for the interim. Nevertheless, we must actually go on to alter our concepts so as to remove the inconsistency.

4.2.3 The Replacement Policy

I advocate the replacement policy for inconsistent concepts. According to it, we should determine the best way to replace an inconsistent concept with consistent ones, and do so. That means we should stop employing an inconsistent concept and begin employing a different concept or group of concepts. One difficult issue with pursuing the replacement policy is the choice of replacement(s). I say very little on how to go about choosing a replacement(s) for an inconsistent concept and I am not sure that it is possible to provide a strategy that will result in the best replacement(s) each time. It seems to me that judging the best replacement(s) involves the weighing of factors that are not easily quantifiable, as in considerations of simplicity, economy, and charity.

I want to address one objection to the replacement policy. It is that an inconsistent concept might be indispensable. That is, it might be so important that we cannot get along without it. Alternatively, it might be so ubiquitous that the replacement policy itself depends on it. It might even turn out to be impossible for any combination of consistent concepts to do the work of the inconsistent one. These three varieties of indispensability claims are problematic for the advocate of the replacement policy. I can offer no guarantee that the replacement policy will be successful for every inconsistent concept. Indispensable inconsistent concepts, if such monsters exist, would be rather troubling and I am afraid that I have nothing helpful to say on the matter. However, the mere possibility of indispensable inconsistent concepts does not undermine the replacement policy or offer a justification for the containment policy as an alternative. The containment policy would be an acceptable alternative to the replacement policy only if good reasons were marshaled for the claim that a particular concept is both inconsistent and indispensable. Moreover, the replacement policy would still be the favored one in the general case. The containment policy would only serve as a fallback position for particularly recalcitrant cases.

4.4 Inconsistent Concepts are Confused

So far in this chapter I have given a definition of ‘inconsistent concept’, introduced several types of inconsistent concepts and discussed three policies for handling conceptual inconsistency. In this section, I advocate a particular explanation of inconsistent concepts. On my view, *inconsistent concepts are confused concepts*. To carry out this explanatory strategy, I owe an account of confused concepts. In the remainder of this chapter I discuss confused concepts and provide some examples. In Chapter Five, I endorse Joseph Camp’s theory of confusion and his logic for confused singular terms. In Chapter Six, I extend Camp’s theory of confusion to concepts and I use it to construct a semantic theory for confused concepts. Chapter Six also includes the full explanation of inconsistent concepts in terms of confused concepts.

It is perhaps easiest to say what it is for a person to be confused. Roughly, a person is *confused* if and only if he or she thinks that two or more distinct entities are identical.[16] The entities in question can be objects, properties, relations, concepts, propositions, etc. One can then define a *confused singular term* as one that is used by a person who thinks that two or more distinct objects are identical, where that person uses the singular term in an attempt to refer to the one object she thinks exists. In the next chapter I discuss an example of Camp’s where Fred is a person who owns an ant farm and mistakenly believes that it contains only one big ant when in fact it houses two big ants. Fred uses the name ‘Charlie’ in an attempt to refer to what he thinks is the sole big ant in the ant farm. In this example, Fred is confused because he thinks that two distinct objects (the two big ants) are identical. ‘Charlie’ as used by Fred is a confused singular term because a confused person uses it in an attempt to refer to the one object he thinks exists.

[16] In this formulation ‘thinks that’ is not synonymous with ‘believes that’. In the next chapter I discuss this use of ‘thinks that’. For now I leave it at an intuitive level.

If we assume that predicates are used to express concepts and to designate properties, then we can define a *confused predicate* as one that is used by a person who thinks that two or more distinct properties are identical, where that person uses the predicate to designate the one property he thinks exists.[17] A *confused concept* is one that is expressed by a confused predicate. The *components* of a confused concept are the distinct concepts expressed by the predicates that designate the properties thought to be identical by the confused person in question. I sometimes say that a confused concept is a *fusion* or an *amalgam* of its components.

An example of conceptual confusion popularized by Field is the Newtonian concept of mass. In Newton's physics, physical objects have a single physical quantity: mass. According to this theory, mass obeys the two laws (which are considered equally fundamental): (i) mass = momentum / velocity, and (ii) the mass of an object is the same in all reference frames. In Einstein's physics, physical objects have two different "kinds" of mass: proper mass and relativistic mass. An object's proper mass is its total energy divided by the square of the speed of light; an object's *relativistic mass* is its non-kinetic energy divided by the square of the speed of light. Although relativistic mass = momentum / velocity, the relativistic mass of an object is not the same in all reference frames. Contrariwise, proper mass ≠ momentum / velocity, but the proper mass of an object is the same in all reference frames. Thus, relativistic mass obeys one of the laws for the Newtonian concept of mass and proper mass obeys the other.

Field argues that the Newtonian concept of mass is a confused concept (although he does not use the term 'confused'). That is, an employer of **mass** thinks that two distinct physical quantities (relativistic mass and proper mass) are identical. Field argues that 'mass' as used by the Newtonian physicist did not designate relativistic mass and did not designate proper mass; it

[17] This use of 'designate' can be found in those who follow Kripke in calling some predicates 'rigid designators' (see Kripke 1972 and Soames 2002). I use it out of convenience; one could instead define confused predicates in terms of extensions.

did not designate some other quantity and it did not fail to designate anything at all. Instead, Field claims that it partially designated both relativistic mass and proper mass.[18] In the next chapter I take up the issue of how best to interpret confused concepts like this one.

According to my theory of inconsistent concepts, an inconsistent concept is a confused concept.[19] Thus, an employer of an inconsistent concept is confused. Hence, an employer of an inconsistent concept that is expressed by a predicate thinks that two or more distinct properties are identical and uses a confused predicate in an attempt to designate the one property he or she thinks exists. That explanation imposes conditions on attributions of conceptual inconsistency. For instance, the claim that a certain concept is inconsistent implies that it is a fusion of two or more concepts. This account dovetails with the replacement strategy for dealing with inconsistent concepts because the component concepts for an inconsistent concept are usually ideal candidates for the replacements.

4.5 Conclusion

In this chapter, I presented an account of inconsistent concepts, on which an inconsistent concept is one whose constitutive rules for employment are incompatible. I defined several types of inconsistent concepts and discussed policies for handling them. Of particular importance is the distinction between application-inconsistent and range-inconsistent concepts, and the distinction

[18] See Field (1973, 1974).

[19] Notice that not all confused concepts are inconsistent. For example, let A be a concept that applies to objects α and β and disapplies to objects γ and δ. One can think of A as a confused concept whose components are B, which applies to α and disapplies to γ, and C, which applies to β and disapplies to δ. A is neither application-inconsistent nor range-inconsistent.

between empirically inconsistent and intrinsically inconsistent concepts. I also proposed to explain inconsistent concepts in terms of confused concepts.

I have not argued for the claim that inconsistent concepts are confused concepts and I do not intend to. That claim is the basis for my theory of inconsistent concepts and I justify my theory by appeal to how well it works. In Appendix E, I consider several different theories of inconsistent concepts and I argue that mine is preferable to all the competitors.

To my knowledge, no one has advocated an explanation of inconsistent concepts in terms of confusion, but we do find some connection between them in the literature. In particular, confused concepts and inconsistent concepts have provoked similar worries and they have been explained in similar ways. For example, Field advocates a semantic theory based on supervaluations for confused concepts in the paper from which the mass example is taken. However, Field does not actually use the term ‘confusion’ to describe the phenomenon. Instead, he talks about referential indeterminacy. For Field, referential indeterminacy is a broader category intended to capture vagueness as well. Supervaluations have been used in semantic theories for truth as well. Indeed, Eklund argues that truth is an inconsistent concept and he uses a variant of the supervaluation approach in his semantic theory for truth. Furthermore, Field argues that confused concepts pose a serious threat to certain theories of meaning, while Gupta and Eklund (independently) argue that inconsistent concepts pose similar threats to theories of meaning. The novelty in my approach consists in the direct link between confusion and inconsistent concepts and the extension of Camp’s theory of confusion and his logic for confused names to a semantic theory for confused concepts.

5.0 A THEORY OF CONFUSION

5.1 INTRODUCTION

I claim that truth is an inconsistent concept. In the preceding chapter, I discussed inconsistent concepts and committed myself to explaining them in terms of confused concepts. This chapter contains my preferred theory of confusion, which Joseph Camp recently presented. In the first section of this chapter, I discuss several of Camp's arguments for conditions on theories of confusion. In the second section, I explain the logic he advocates for confused names (although he really defines it only for confused sentences). In Chapter Six, I extend Camp's theory in several ways and complete my explanation of inconsistent concepts in terms of confusion.

5.2 CAMP'S THEORY OF CONFUSION

Consider a person, Fred, who buys an ant farm and dumps some ants into it. Although two large ants fall into the ant farm, Fred sees only one of them go in. Fred does not know that there are two ants in the ant farm and, due to some fact about large ant behavior, they are never visible together. One day when Fred is away from his ant farm he decides to use 'Charlie' as a name for what he takes to be the only big ant in the ant farm. (The condition that he is not in close proximity to the ant farm combats the temptation to assume that 'Charlie' refers to whichever big ant Fred perceived when he coined the name.) Fred routinely studies the two big ants and uses the name 'Charlie' to keep track of his findings; but he never discovers that there are two big ants in the farm. To help clarify matters, I use the names 'Ant A' and 'Ant B' for the two big ants in the ant farm. One can characterize Fred's confusion by saying that Fred has confused Ant A with Ant B, or that Fred thinks that Ant A is Ant B.[1]

A theory of confusion has implications for the inferential rationality of confused people. In other words, when one adopts a certain theory of confusion one decides how to treat a confused person's reasoning practice. For instance, a theory of confusion should specify a logic for arguments that contain confused expressions. A *logic* is a theory that specifies which of the arguments in some class are valid. Thus, when one adopts a theory of confusion, one undertakes a commitment to evaluate confused arguments according to a certain standard and to treat confused people as if they should reason according to that standard.

Perhaps the most fundamental of Camp's assumptions is that a theory of confusion should be inferentially charitable. It should not imply that confused people are poor reasoners. Prior to his acquisition of the ant farm, Fred knew how to construct valid arguments and how to evaluate arguments for validity (we can even assume that Fred is an eminent logician). He knew how to follow inference rules and how to weigh evidence for and against a claim. The acquisition of the ant farm does not change these facts. If a theory of confusion implies that Fred no longer has these abilities after his acquisition of the ant farm, then that theory is false. This inferential charity requirement on theories of confusion is at the heart of several of Camp's arguments for four further conditions on theories of confusion: confusion should not be explained in terms of ambiguity, confusion should not be explained in terms of belief, sentences with confused expressions do not have truth-values, and a logic for confused expressions ought

[1] Camp (2002: 27-29).

to be one that the confused person has a reason to obey. I consider each of these conditions in the next four subsections.

5.2.1 The Ambiguity Condition

Camp distinguishes between confusion and ambiguity, and refuses to explain the former in terms of the latter. An *ambiguous expression* is one that has more than one distinct semantic value. 'Semantic value' is a generic term for the contribution an expression makes to the meaning of the sentences in which it occurs (e.g., reference, meaning, etc.). A typical example of an ambiguous word is the English word 'bank'. On some occasions it is synonymous with 'side of a river', while on others it has a meaning similar to that of 'financial institution'. Someone who explains confusion in terms of ambiguity holds that a confused expression has more than one semantic value. In the case of Fred, the most natural explanation of this sort is that sometimes 'Charlie' refers to Ant A and other times it refers to Ant B. An explanation of confusion in terms of ambiguity should specify a disambiguation rule, which determines the referent of any given occurrence of 'Charlie'. Presumably, this theory of confusion would also use a classical logic to evaluate Fred's confused arguments (so long as such a logic is appropriate for Fred's other arguments).

Any explanation of confusion in terms of ambiguity will treat some of Fred's arguments as equivocations. An *equivocation* is an invalid argument that contains multiple occurrences of an ambiguous expression where some occurrences have one semantic value and others have another one. For example, Fred presents the following argument:

(i) Charlie is angry.

(ii) Charlie is hungry.

(iii) Charlie is an ant.

(iv) If an ant is angry and hungry, then it is dangerous.

∴ (v) Charlie is dangerous.

If the rule for disambiguating ‘Charlie’ specifies that in (i) ‘Charlie’ refers to Ant A and in (ii) ‘Charlie’ refers to Ant B, then this argument is invalid because it is an equivocation. Because Fred does not know that ‘Charlie’ is ambiguous, he will invariably assert arguments that count as equivocations no matter what rule is used to disambiguate ‘Charlie’.

A theory of confusion that treats Fred as if he endorses equivocal arguments is not inferentially charitable. Equivocations are serious fallacies and someone who commits them is a poor reasoner. However, acquiring the ant farm should not affect Fred’s capacity to reason properly. Thus, a theory of confusion should not explain confusion in terms of ambiguity.[2, 3]

5.2.2 The Belief Condition

Although it might be tempting to claim that Fred counts as confused because he holds the false belief that there is exactly one big ant in the ant farm or that he holds the false belief that Ant A is identical to Ant B, Camp argues that confusion should not be explained in terms of belief. Camp’s argument divides into two cases: confusion as *de dicto* belief and confusion as *de re* belief. Camp’s argument that confusion is not a matter of *de dicto* belief is that Fred might hold any particular *de dicto* beliefs for reasons that are independent of his interactions with Ant A and

[2] Camp considers the objection that a defender of the ambiguity account could specify that the confused person should not be held responsible for the equivocations. Camp’s reply is that calling someone’s argument an equivocation but refusing to hold that person responsible for the error is not being inferentially charitable—it is treating the person as if he is a poor reasoner and then treating him as if he is too stupid to be held accountable for his mistakes, which is even worse. See Camp (2002: 50-54).

[3] It seems to me that this argument compliments Kripke’s claim that positing ambiguity is the lazy person’s way to do philosophy. Not only is it lazy, it is uncharitable. See Kripke (1977: 19).

Ant B. Thus, while having certain beliefs might count as a necessary condition for confusion, it is not sufficient. For example, one might hold that Fred is confused if and only if he believes that there is only one big ant in the ant farm. However, it is possible that Fred believes that there is only one big ant in the ant farm without being confused (for example, if he pays no attention to how many big ants are in it and someone he trusts tells him that there is only one big ant in it). A similar criticism holds for the claim that Fred is confused if and only if he believes that the ant that he saw at 1pm is identical to the ant he saw at 2pm (or any other identity claim of this type). Camp's argument that confusion is not a matter of *de re* belief is based on the claim that attributing to Fred the *de re* belief that Ant A is identical to Ant B represents Fred as someone who can distinguish between Ant A and Ant B, which is false. Thus, Fred has no such *de re* beliefs.[4]

I endorse both these arguments but in my experience, some people find them unconvincing. Instead of reassuring the reader, I want to present a different argument that Camp hints at but does not give. Confusion should not be explained in terms of a false belief or false set of beliefs because, as I argue in the next subsection, the sentences a confused person uses to express his beliefs have no truth-values. Thus, the beliefs expressed by these sentences have no truth-values. Hence, they are not false (that 'not' expresses exclusion negation). Therefore, confusion should not be explained in terms of false beliefs. This argument assumes that the beliefs expressed by confused sentences (i.e., those sentences that contain confused expressions) are the ones that would be used as a basis for the explanation of confusion and that if a sentence has no truth-value then the belief it expresses has no truth-value either. These assumptions are plausible enough that I do not want to linger over them any further. Instead, I move directly to the main assumption of this argument—that confused sentences have no truth-values.

[4] A similar argument can be found in Dennett (1981).

5.2.3 THE TRUTH-VALUE CONDITION

The third condition on theories of confusion is that confused sentences have no truth-values. To make his case for this condition, Camp uses the notion of calibration. A person is *well calibrated* with respect to some event if the subjective probability she assigns to the event is close to its objective probability. The argument is that if one assigns truth-values to Fred's confused sentences, then one treats Fred as if he is poorly calibrated when he is not. Calibration is largely a matter of being able to weigh evidence for and against some claim and make judgments accordingly. A poorly calibrated person is not good at inductive reasoning on the topic in question. Thus, if confused sentences have truth-values, then the confused person is poorly calibrated, and a poorly calibrated person is a poor inductive reasoner. Therefore, assigning truth-values to confused sentences is inductively uncharitable.

Consider Camp's example in which Ant A has just come down with the sniffles and Ant B has had them for three days. Fred knows that sniffling ants continue to sniffle for an average of 2.63 days. He inspects Ant B and discovers that it has the sniffles. Fred endorses the following argument:

(i) Charlie has the sniffles.

∴ (ii) Charlie will continue to sniffle for an average of 2.63 days.

Both ants have the sniffles so if one assigns truth-values on the basis of supervalutions, the premise is true but the conclusion is false. Or at least, if one treats it as true, an utterance of the same token will be false on average, in less than 2.63 days. However, just as with his deductive arguments, Fred is good at reasoning. He is just confused.[5]

I endorse the calibration argument, but I want to present another argument for the same conclusion based on the claim that confused expressions are defective. Some remarks of Dummett's on concepts will serve to make a point about defective concepts. In the passage below, Dummett comments on the distinction between the circumstances of application and the consequences of application for a given concept (an account of concepts that employs this distinction is analogous to Gentzen's account of logical connectives that employs introduction and elimination rules).[6]

> The distinction is thus meant as no more than a rough and ready one, whose application, in a given case, will depend in part on how we choose to slice things up. It remains, nevertheless, a distinction of great importance, which is crucial to many forms of linguistic change, of the kind we should characterize as involving the rejection or revision of concepts. Such change is motivated by a desire to attain or preserve a harmony between the two aspects of an expression's meaning. A simple case would be that of a pejorative term, e.g., 'Boche'. The condition for applying the term to someone is that he is of German nationality; the consequences of its application are that he is barbarous and more prone to cruelty than other Europeans. … Someone who rejects the word does so because he does not want to permit a transition from the grounds for applying the term to the consequences of doing so, (Dummett 1973: 454).

One important feature of Dummett's model is that it permits a characterization of defective concepts like **Boche**. I possess this concept; however, I do not employ it. I do not employ this

[5] Camp hints at a different argument that is more epistemological in nature. He claims that truth attributions have epistemological consequences. If Uter asserts 'Fred's assertion that Charlie is asleep is true', and Adil hears him, then Adil will assume that Uter believes the same thing Fred believes. If Uter knows that Fred is confused, and he cares about his own cognitive well being, then he should not accept Fred's belief, nor should he defer to Fred. Uter knows that Fred is untrustworthy with respect to 'Charlie' talk. Hence, he should not attribute truth to Fred's confused sentences. I am less sympathetic to this argument. Camp's point seems to be that if confused sentences have truth-values, then surely we should be able to say so. However, when we attribute truth-values to them, our attributions carry dubious epistemological considerations in their wake. Hence, we should not attribute truth-values to confused sentences. Thus, they do not have truth-values. It seems to me that the epistemological consequences of truth attributions might well be a case of conversational implicature. If so then it might be perfectly fine to attribute truth-values to confused sentences so long as one is careful to prevent the implicature.

[6] See Gentzen (1969). See also Brandom (1994: 116-130) for a similar account of concepts.

concept because I disagree with it in some sense. In particular, I reject the inference from its conditions of application to its consequences of application. That is, I do not endorse the inference from 'α is of German ancestry' to 'α is barbarous and more prone to cruelty than other Europeans'. Even if α is German, it is inappropriate to assert 'α is Boche' because it follows that α is barbarous and more prone to cruelty than other Europeans. Likewise, it is inappropriate to assert 'α is not Boche' because it follows that α is not of German descent. Thus, it is inappropriate to employ **Boche**. However, I certainly possess it. I can attribute it to others and I understand claims made with it. The same distinction is important for confused concepts. Without the ability to distinguish between concept *employment* and concept *possession*, it is impossible to give a plausible account of how one person can attribute a confused concept to another without falling into confusion herself. See Appendix E for more discussion of this issue.

If it is inappropriate to employ a defective concept like **Boche**, then it is inappropriate to attribute truth-values to sentences that express it. If it is inappropriate to assert 'α is Boche', then it is inappropriate to assert ''α is Boche' is true', because the former follows from the latter on any account of truth that is remotely plausible. Likewise, it is inappropriate to assert ''α is Boche' is false' because 'α is not Boche' follows from it. If it is inappropriate to attribute truth values to sentences that express defective concepts, then these sentences do not have truth-values. (These points hold for beliefs whose content involves defective concepts as well.) I suppose that one could endorse the view that such sentences secretly have truth-values but we cannot know what they are and cannot find out, but this sort of view is implausible. As I use the term 'truth-value', if a sentence has a particular truth-value, then it is appropriate to say and believe that it does.

I still need to show that confused concepts are defective in the same way that **Boche** is. The easiest way to see this is to consider the example from the previous chapter of **mass**. If I assert that a particular object has mass then I am committed to the claim that its momentum divided by its velocity is the same in all reference frames, which is false. If I assert that that object has no mass, then I am committed to the claim that it has no momentum in any reference frame, which is false as well. Thus, it is inappropriate for me to assert of any object that it has mass or that it does not have mass. That is, it is inappropriate for me to employ **mass**. **Mass** is a defective concept. Thus, sentences that express it have no truth-value.

It is important to realize that my view is not that whether a sentence has a truth-value is relative to a person or group of people. Truth-values and defectiveness are objective on my account. However, we can be wrong about which sentences have truth values and which ones do not, just as we can be wrong about which sentences are true and which are false. If confused sentences have truth-values, then it is appropriate for even unconfused people to attribute truth-values to them. If it is appropriate to attribute truth-values to them, then it is inappropriate for someone to refuse to employ the confused concept in question. However, I have argued that it is appropriate to refuse to employ confused concepts. Hence, confused sentences have no truth-values.

5.2.4 The Normativity Condition

Here is Camp's formulation of the normativity condition (not his term) on the logic for confused expressions that follows from a theory of confusion: "Roughly, the requirement is that the person must be moved by the fact that the conclusion of an argument does or does not follow from the premises. The person must care, in principle, whether an argument is valid," (Camp 2002: 79).

This might seem like an odd constraint on a logic. After all, the validity of an argument does not depend on whether the person who presents it can tell whether it is valid. It seems like the mental state of the person in question is irrelevant to the validity of his or her arguments. Why should the validity of a person's arguments be held hostage to what the person cares about? Camp does not offer much in the way of argument for the normativity condition, but does give us a clue as to why someone might hold it: "A similar requirement applies to many other norms of practice; there is nothing special about logical norms here. For example, to be moral a person must do what is right (mostly), and must care, in principle, whether a prospective action is right," (ibid). I want to spell out what I take to be the bigger issues behind the normativity condition so that it does not seem so counterintuitive.

The first point I want to make is that when one formulates a validity criterion for the arguments that display some vocabulary, one specifies how rational agents who use that vocabulary *should* reason. It is a specification of the inferential norms that have authority over those rational agents. In particular, a theory of confusion that specifies a certain logic for confused arguments is a characterization of the inferential rationality of the confused. The confused are not irrational. If a theory of confusion entails that a confused person is irrational, then that theory is false. That is the problem (in general terms) with the claim that confused expressions are ambiguous and with the claim that confused sentences have truth-values—these claims entail that confused people are irrational.

There is a view of normativity, the Kantian conception, according to which the authority that norms have over rational agents derives from the fact that rational agents accept or endorse those norms in a certain way. The debates about normative authority take place primarily over the source of practical normativity, but the same points can be made for theoretical contexts.

The following is a passage in which Robert Brandom endorses the Kantian conception as a general constraint on accounts of rationality:

> [Kant] characterizes [normative compulsion] substantively as acting according to a *conception* or *representation* of a rule, rather than just according to a rule. Shorn of the details of his story about the nature of representations and the way they can affect what we do, the point he is making is that we act according to our *grasp* or *understanding* of rules. The rules do not immediately compel us, as natural ones do. Their compulsion is rather mediated by our *attitude* toward those rules. What makes us act as we do is not the rule or norm itself but our *acknowledgement* of it, (Brandom 1994: 31).[7]

With respect to a theory of confusion, the logic for confused expressions that is part of such a theory is a specification of the inferential norms that bind those who employ confused expressions. If a theory of confusion entails that someone who employs a confused expression has no reason to adopt the inferential norms specified by that theory, then the person who employs the confused expression is irrational according to that theory. This is not the place for a detailed discussion and defense of the Kantian conception of normativity. However, identifying it as one possible basis for Camp's normativity constraint should eliminate much of the mystery surrounding the constraint.

Another doctrine familiar in debates over practical rationality is practical internalism. *Practical internalism* is the view that there is a necessary connection between practical reasons and motivation.[8] Here is a formulation of practical internalism due to Setiya: "If the fact that *p* is a reason for A to ϕ then: if A is not disposed to be moved to ϕ if she judges that *p*, A is in that respect *practically irrational*," (Setiya 2004: 268). If we transplant practical internalism from

[7] See Korsgaard (1996) for an example of this position in meta-ethics and Harman (1986, 1995) for discussions of these issues as they arise in theoretical contexts.

[8] I modify 'internalism' with 'practical' to distinguish the doctrine under consideration from other types of internalism (e.g., *epistemological internalism*—the doctrine that if a belief counts as knowledge then the believer must know that it does—and *semantic internalism*—the doctrine that the meanings of one's terms and the contents of one's mental states are independent of the physical and social environment one inhabits). There are other doctrines that go by the name 'internalism' and pertain to practical rationality. I do not have the space to discuss them here.

the context of practical rationality into our context (inferential rationality and the relation between a theory of confusion and the capacity of the confused to care whether a confused argument is valid according to that theory), then we arrive at the following doctrine. If the fact that an inferential norm is valid is a reason for a confused person to reason in accordance with that norm then: if the confused person is not disposed to be moved to reason in accordance with that norm if she judges that that norm is valid, the confused person is in that respect irrational.[9] In other words, if a confused person recognizes that a certain inferential norm is valid but is not motivated to infer according to it, then that person is irrational. From this doctrine, we can derive the normativity constraint. If a theory of confusion specifies a certain logic for confused expressions and someone who employs a confused expression can recognize when arguments are valid according to that logic, but that person is not motivated to infer according to that logic, then that person is irrational according to that theory of confusion. However, confused people are not irrational. Thus, a theory of confusion with this property is false.

I do not take a stand on whether it is the Kantian conception of normativity or a version of practical internalism that motivates Camp's normativity constraint (although there are important connections between the Kantian conception and practical internalism, they are distinct doctrines). Nor do I attempt to defend one of these doctrines. I do accept the normativity constraint on theories of confusion and I accept it because I accept a Kantian conception of normativity. Given that either the Kantian conception of normativity or a version of practical internalism will justify the normativity constraint, the reader should feel free to think of this disjunction as one of my assumptions.

[9] Setiya goes on to criticize a version of practical internalism, but that criticism does not apply to the version of the doctrine I formulated because Setiya's criticism applies to practical internalism for objective reasons and my version appeals to subjective reasons. A formulation of practical internalism for inferential rationality that appeals to objective reasons is not plausible in the case of confusion. See Setiya (2004: 271-272).

5.2.5 SEMANTIC STANCES

Camp explains confusion in terms of adopting a certain semantic stance.[10] When a person (Ginger, for example) utters ‘Fred thinks that Ant A is Ant B’, she is not attributing some mental state to Fred. Instead, she adopts a semantic stance toward Fred and his confused sentences. When Ginger adopts a semantic stance toward Fred, she alters what she counts as a correct inference; she decides to be inferentially charitable to Fred in a certain way. I call the particular stance that one should adopt when interpreting a confused person the *confusion stance*. According to Camp’s theory of confusion, a person is confused if and only if it is appropriate to adopt the confusion stance toward that person. By characterizing the confusion stance, we arrive at a logic for confused expressions.

Camp’s theory of confusion explains what it is to be confused in terms of what it is to treat someone as confused. That theory implies that confusion is a status that a person, concept, expression, sentence, or argument can have. I want to emphasize that his theory employs a particular explanatory strategy; it does *not* imply that a concept is confused *because* someone takes it to be confused; it does not imply that confusion is relative to an interpreter; it does not imply that confusion is not “real” or objective. Explaining a status in terms of attributing that status or treating something as if it has that status is compatible with the claim that something can have that status despite the fact that no one treats it that way.

Camp is silent on the issues of what it is to adopt a semantic stance and when it is appropriate to adopt the confusion stance in particular. In the next chapter I address the first

[10] Camp uses the term ‘semantic position’ instead. It seems to me that the notion of a stance is more familiar. See Dennett (1987) and Brandom (1994: 55-64; 2002: 1-17).

issue, but I have nothing to say on the second. How we (humans) actually decide to treat something as confused is an issue for psychology, not a theory of confusion.

5.3 Camp's Logic for Confused Sentences

Camp advocates a particular logic by which one should evaluate the inferences of the confused. Because Camp argues that confused sentences have no truth-values, he must present an inferential standard by which one can evaluate a confused person's inferences that does not define validity in terms of truth preservation. Instead, he defines validity in terms of profitability preservation. A sentence is *profitable* if and only if believing it will contribute to the achievement of one of the believer's goals.[11] Despite the fact that he is confused, Fred's beliefs still have the same causal role in producing his actions and some of his confused beliefs will be more profitable than others in the sense that acting on a profitable one is more likely to satisfy Fred's desires than acting on an unprofitable one. (For example, if both ants are angry and Fred desires that he stay away from Charlie when Charlie is angry, then Fred's belief that Charlie is angry is more profitable than the belief that Charlie is not angry—despite the fact that neither of these beliefs have truth-values.) We can say that a confused argument is *valid* if and only if it preserves profitability.

Camp uses Belnap's useful four-valued logic to track profitability.[12] This logic uses four semantic values: Y, N, ?, and YN. Camp uses a story about two people, Sal and Sam, who are authorities on the properties of the ant farm (e.g., they are not confused about Ant A and Ant B),

[11] Camp (2002: 122-124).

[12] See Belnap (1976a, 1976b) and Dunn (1966). Belnap advocates using the logic for the inferences one should draw from a database with inconsistent information.

to illustrate the intended interpretation of these semantic values. The idea is that their opinions are indicators of profitability for Fred. Let us rejoin Ginger in her attempt to find a way to be inferentially charitable to Fred. Ginger should begin by assigning semantic values to Fred's sentences in the following way. Assume that Fred utters a confused sentence, *p*, with the term 'Charlie' in it. Ginger should substitute 'Ant A' for 'Charlie' and ask Sal to evaluate the resulting sentence. If Sal agrees with the resulting sentence, he says, "Yes"; if not, he says, "No." If he does not have enough information on which to evaluate the sentence, he says, "I don't know." Ginger should substitute 'Ant B' for 'Charlie' in Fred's sentence and ask Sam to evaluate it as well. If both Sal and Sam say, "Yes" or one says, "Yes" and the other says, "I don't know," then Ginger should assign Y to Fred's sentence. If both say, "No" or one says, "No" and the other says, "I don't know," then Ginger should assign N to Fred's sentence. If both say, "I don't know," then Ginger should assign ? to Fred's sentence. If one says, "Yes" and the other says, "No," then Ginger should assign YN to Fred's sentence.

The semantic values are grouped as follows: if a sentence is Y or YN, then it is *at-least-Y* and if a sentence is N or YN, then it is *at-least-N* (we can say that Y and YN are *designated*). Ginger can now use the following standard to evaluate Fred's arguments: an argument is valid just in case it preserves at-least-Y (i.e., if the premises are at-least-Y, then the conclusion is at-least-Y) and the absence of at-least-N (i.e. if the conclusion is at-least-N, then one of the premises is at-least-N).

We have a way of assigning semantic values to Fred's confused sentences, we have an interpretation of the semantic values, and we have a validity criterion based on the interpretation of the semantic values. We still need to introduce logical connectives.[13] The following are the tables for negation (~), conjunction (∧), and disjunction (∨).[14]

p	~ p
?	?
Y	N
YN	YN
N	Y

∧	?	Y	YN	N
?	?	?	N	N
Y	?	Y	YN	N
YN	N	YN	YN	N
N	N	N	N	N

∨	?	Y	YN	N
?	?	Y	Y	?
Y	Y	Y	Y	Y
YN	Y	Y	YN	YN
N	?	Y	YN	N

These tables are constructed according to the following rules:

⟨~ p⟩ is at-least-Y if and only if ⟨p⟩ is at-least-Y.

⟨~ p⟩ is at-least-N if and only if ⟨p⟩ is at-least-N.

⟨p ∧ q⟩ is at-least-Y if and only if both ⟨p⟩ and ⟨q⟩ are at-least-Y.

⟨p ∧ q⟩ is at-least-N if and only if either ⟨p⟩ or ⟨q⟩ are at-least-N.

⟨p ∨ q⟩ is at-least-Y if and only if either ⟨p⟩ or ⟨q⟩ is at-least-N.

⟨p ∨ q⟩ is at-least-N if and only if both ⟨p⟩ and ⟨q⟩ are at-least-N.

Given that at-least-Y is designated, these rules are extensions of the classical rules for negation, conjunction, and disjunction.

[13] We cannot allow the semantic value assignment and the validity criterion to do all the work. Consider an example. Ant A is hungry but not sleepy and Ant B is sleepy but not hungry. Fred argues: Charlie is hungry; Charlie is sleepy; hence, Charlie is sleepy and hungry. The value assignment for this argument will be YN, YN and N, respectively. Thus, the argument is invalid according to the validity criterion. However, this argument is an instance of conjunction introduction, which should be valid. Thus, as far as the logical constants of Fred's language go, we need to take them out of the hands of Sam and Sal.

[14] Camp (2002: 125-157).

The following diagram displays the inferential relations between the semantic values:

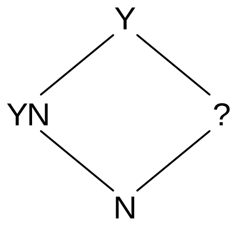


Figure 5.1 (Four Valued Logic)

The lines in the diagram indicate implication. If we think of the semantic values as ordered from bottom to top, and write ‘v(⟨p⟩) ≤ v(⟨q⟩)’ for ‘the semantic value of ⟨p⟩ is less than or equal to the semantic value of ⟨q⟩’ then ⟨p⟩ entails ⟨q⟩ if and only if v(⟨p⟩) ≤ v(⟨q⟩) for every assignment of values to ⟨p⟩ and ⟨q⟩ We can think of conjunction as “meet”; thus, the value of ⟨p ∧ q⟩ is the greatest lower bound of the values of ⟨p⟩ and ⟨q⟩ (e.g., if ⟨p⟩ is YN and ⟨q⟩ is ?, then ⟨p ∧ q⟩ is N, which is the greatest value that is less than or equal to each). We can think of disjunction as “join”; thus, the value of ⟨p ∨ q⟩ is the least upper bound of the values of ⟨p⟩ and ⟨q⟩ (e.g., if ⟨p⟩ is YN and ⟨q⟩ is ?, then ⟨p∨q⟩ is Y, which is the least value that is greater than or equal to each).[15]

So far I have discussed only negation, conjunction, and disjunction. We can add an additional connective ‘→’ such that ⟨p → q⟩ stands for ⟨p entails q⟩. It is only defined for sentences that do not contain ‘→’. Dunn proved that ⟨p → q⟩ is valid in this logic if and only if

[15] We can summarize this paragraph by saying that the lattice of semantic values is a DeMorgan lattice.

$\langle p \rightarrow q \rangle$ is a theorem of the relevance logic R_{fde} (the logic of first-degree implications).[16] An axiomatization of R_{fde} uses the following axioms and rules:[17]

Axioms:
$p \rightarrow \sim\sim p$
$\sim\sim p \rightarrow p$
$(p \wedge q) \rightarrow p$
$(p \wedge q) \rightarrow q$
$p \rightarrow (p \vee q)$
$q \rightarrow (p \vee q)$
$(p \wedge (q \vee r)) \rightarrow ((p \wedge q) \vee r)$

Rules:
$p \rightarrow q, q \rightarrow r \vdash p \rightarrow q$
$p \rightarrow q, p \rightarrow r \vdash p \rightarrow (q \wedge r)$
$p \rightarrow r, q \rightarrow r \vdash (p \vee q) \rightarrow r$
$p \rightarrow q \vdash \sim q \rightarrow \sim p$

The connection between the 4-valued logic presented above and R_{fde} is a significant result. It means that the appropriate logic for evaluating confused arguments is a relevance logic.

Relevance logics are so named because they require that an argument's premises must be relevant to its conclusion for it to count as valid. A related point is that in relevance logics, the antecedent and the consequent of a conditional must be relevant for the conditional to be true. According to classical logic, one can add premises to a valid argument without changing it from valid to invalid even if the premises are not used to derive the conclusion. Likewise, according to the semantics for the material conditional of classical logic, 'if 0=1, then I am president of the United States' is true because the antecedent is false. However, many people find these results counterintuitive because the premises and conclusion in the argument and the antecedent and the consequent of the conditional have nothing to do with one another. There are a variety of relevance logics that differ over how they spell out the connection between premises and

[16] Dunn (1966). See Dunn and Restall (2001: 54). Note that Anderson and Belnap use 'E_{fde}' instead of 'R_{fde}'. I should note that Camp does not explicitly endorse the introduction of '→' into the logic, but he does say that the class of implications in the 4-valued logic are those valid in R_{fde} (Camp 2002: 157). Presumably, he would not object to the introduction of '→'.

[17] See Dunn and Restall (2001: 27-28) and Anderson and Belnap (1975). In the inference rules '$\vdash$' should be read as 'implies'.

conclusion of an argument and the relation between antecedent and consequent of a conditional. One thing that is common to all of them is that they are non-classical. That is, some forms of inference that are valid in classical logic are invalid in relevance logics (e.g., disjunctive syllogism—for any two sentences p and q, ⌜~q⌝ follows from ⌜p or q⌝ and ⌜~p⌝). I discuss relevance logics more in the next chapter.

5.4 Conclusion

My goal in this chapter is to present the theory of confusion I endorse. Although I accept the four conditions on theories of confusion, they are not arguments for the claim that confusion should be explained in terms of adopting a semantic stance or the claim that a relevance logic is the best one for evaluating the arguments of the confused. However, I offer no more justification for Camp's theory of confusion. If one is not convinced that Camp's theory is preferable to other accounts of confusion, then one should replace my claim that inconsistent concepts are confused with the claim that a theory of inconsistent concepts should be based on Camp's theory of confusion. That is, because I offer a justification for the theory of inconsistent concepts I develop, I have no need to justify Camp's theory of confusion. The arguments in section one are intended to give the reader an idea of my reasons for accepting Camp's theory, not as conclusive reasons to accept it.

As a theory of confusion, Camp's theory remains incomplete. First, despite the fact that Camp endorses his theory as an account of confused concepts, his logic is applicable only to confused sentences. If it is to serve as a theory of confused concepts, then it must be able to handle confused concepts whose components are partial concepts. Second, the relevance logic that Camp presents is a sentential logic (i.e., it treats atomic sentences as single units), but it does not even include an embeddable conditional. It needs to be fitted with a suitable conditional and extended to a first-order predicate logic. Third, someone who thinks that three distinct entities are identical is confused as well, but Camp's theory works only for confusions associated with two entities. It ought to be extended so that it can handle confusions involving any finite number of entities. Fourth, Camp presents a logic for confused expressions, but not a semantic theory for confused expressions. That is, his theory can tell us when an argument that contains a confused expression is valid, but it does not specify the meanings of the sentences that contain confused expressions. It should be extended to a full semantic theory for confused expressions. In the next chapter, I take up each of these tasks and apply the extended theory to inconsistent concepts.

6.0 Conceptual Confusion and Conceptual Inconsistency

6.1 Introduction

In this chapter I extend Camp's theory of confusion so that it can serve as the basis for a theory of inconsistent concepts. The resulting theory of confusion is extended from Camp's theory in four ways: it applies to n-component confusion, it applies to partially defined confused concepts, it employs a quantified relevance logic, and it specifies a semantic theory for confused expressions that is based on Brandom's theory of meaning. After presenting these extensions, I consider the consequences of explaining inconsistent concepts in terms of confused concepts and provide replies to some objections.

6.2 First-Order Logic for Confusion

Two things need to get accomplished in this subsection. First, I need to provide a conditional that can be added to the sentential logic R_{fde} I discussed in Chapter Five; that will yield a full sentential logic for confused sentences. Second, I need to present a first-order predicate logic that is a natural extension of the sentential logic; that will yield a first order logic for confused expressions. I take up these two tasks in order.

6.2.1 Conditionals

In Chapter Five, I suggested that we add a conditional to the 4-valued logic Camp advocates for the evaluation of confused arguments. I call this 4-valued logic without a conditional *2-component logic* (this choice of term will become clear in Section 6.3). R_{fde} is the result of adding '→' to 2-component logic. However, this conditional is not very useful because it is not embeddable (i.e., the antecedent and consequent cannot contain '→'). There are two sentential logics that are obvious choices for extending 2-component logic with a conditional: R and E. Both R and E are relevance logics. R is the logic of relevant implication (as opposed to material implication, which is found in classical logic), while E is the logic of entailment. The traditional story is that entailment differs from implication in that implication is a relation between a set of sentences (premises) and a sentence (conclusion) that can hold contingently, whereas entailment is just like implication except that entailment holds necessarily. This explanation might lead one to assume that entailment is necessary relevant implication. In other words, it might seem that we could add a necessity operator (□) to R (the logic of relevant implication) and arrive at E (the logic of entailment). Unfortunately, it turns out that R with a necessity operator (called *NR*) has a theorem that is not a theorem of E. Nevertheless, there is an important sense in which E is stronger than R.[1] For this reason, I suggest that an expression for relevant implication be added to 2-component logic to arrive at a 4-valued logic that is equivalent to R. Because the extended 4-valued logic and R are equivalent, it does not matter whether one shows how to add a conditional to 2-component logic to arrive at the extended 4-valued logic or to R_{fde} to arrive at R. Below is an axiomatixation of R:[2]

[1] See Mares (2004) for details.

[2] This axiomatization is due to Mares; see Mares (2004: 208-9).

Axioms: $p \rightarrow p$
$(p \rightarrow q) \rightarrow ((q \rightarrow r) \rightarrow (p \rightarrow r))$
$p \rightarrow ((p \rightarrow q) \rightarrow q)$
$(p \rightarrow p \rightarrow q)) \rightarrow (p \rightarrow q)$
$(p \wedge q) \rightarrow p$
$(p \wedge q) \rightarrow q$
$p \rightarrow (p \vee q)$
$q \rightarrow (p \vee q)$
$((p \rightarrow q) \wedge (p \rightarrow r)) \rightarrow (p \rightarrow (q \wedge r))$
$(p \wedge (q \vee r)) \rightarrow ((p \wedge q) \vee (p \wedge r))$
$\sim \sim p \rightarrow p$
$(p \rightarrow \sim q) \rightarrow (q \rightarrow \sim p)$

Rules: $p, q \vdash p \wedge q$
$p \rightarrow q, p \vdash q$

I assume that if one had a reason to prefer E to R, then one could add an entailment connective to R_{fde} and use E to evaluate confused arguments for validity because the first-degree fragment of R is equivalent to the first-degree fragment of E.[3]

6.2.2 Quantifiers

A sentential logic for confused arguments is an important step, but much of the reasoning that humans use involves quantifiers. To extend our sentential logic to a first-order logic, we have two choices. First, we can develop a predicate logic that is an extension of the 2-component logic (the one with 4 values) or we can develop a predicate logic that is an extension of R. These two options might turn out to be different. I present both in this subsection.

The first-order extension of R is called RQ. Assume that we have the usual language with quantifiers (universal and existential), n-place predicates, individual constants, variables, and the sentential connectives of R. Assume that the elements of the language are assigned

[3] See Anderson and Belnap (1975).

members and subsets of members of a domain D in the usual way. The following is an axiomatization of RQ:[4]

Axioms: all the axioms of R

$\forall xA \to A(c/x)$

$\forall x(A \to B) \to (\forall x A \to \forall xB)$

$\forall x(A \vee B) \to (A \vee \forall xB)$, where x is not free in A

Rule: $A \vdash \forall xA$

The semantics for RQ can be found in Fine (1988).

To extend 2-component logic with relevant implication to first-order 2-component logic with relevant implication, assume that we have a language with quantifiers (universal and existential), n-place predicates, individual constants, variables, and the sentential connectives of 2-component logic with relevant implication. Define a domain D of objects and assign each individual constant an member of the domain and each n-place predicate a pair of subsets of D^n (D^n is the set of n-tuples of elements of D) such that these subsets jointly exhaust D^n.[5] One of the subsets assigned to a predicate P is the extension of P and the other is the antiextension of P. An atomic sentence of the form $\langle Pa_1 \ldots a_n \rangle$ is at-least-Y if and only if the set of n-tuples $<b_1, \ldots, b_n>$ is a member of the extension of P where b_i is the member of D assigned to a_i. Likewise, an atomic sentence of that form is at-least-N if and only if the relevant set of n-tuples is a member of the antiextension of P.

The connectives are handled just as they are in the sentential logic. We define the quantifiers so that universal quantification is analogous to infinite conjunction and existential quantification is analogous to infinite disjunction.

[4] Again based on Mares (2004: 214).

[5] The last assumption might seem odd in this context because most interpretations of 4-valued logics like 2-component logic take ? to act like a gap (neither true nor false). However, on our interpretation, ? acts like a lack of information. In section four, I introduce a family of partial n-component logics that allows for gaps.

$\langle\forall x\phi x\rangle$ is at-least-Y if and only if for every member α of D, the result of substituting the name of α for 'x' in $\langle\phi x\rangle$ is at-least-Y

$\langle\forall x\phi x\rangle$ is at-least-N if and only if for some member α of D, the result of substituting the name of α for x in $\langle\phi x\rangle$ is at-least-N.

$\langle\exists x\phi x\rangle$ is at-least-Y if and only if for some member α of D, the result of substituting the name of α for 'x' in $\langle\phi x\rangle$ is at-least-Y

$\langle\exists x\phi x\rangle$ is at-least-N if and only if for every member α of D, the result of substituting the name of α for 'x' in $\langle\phi x\rangle$ is at-least-N.

Given that at-least-Y is the designated value, these quantifier rules are extensions of the classical rules for quantified sentences. I call this logic *first-order 2-component logic*.

I will not speculate on the relation between first-order 2-component logic and RQ. If they are not equivalent, then further work needs to be done to determine which one is more appropriate for use as a first-order logic for confused arguments.

6.3 N-Component Confusion

Up to this point, I have discussed only 2-component logic. We can think of this logic as one whose semantic values have at most two components (out of Y, N and ?). We can also think of it as the logic that is appropriate for confusion in which two things are thought to be identical. In this section, I introduce a family of logics that can be called *n-component logics*.

Let us say that a confused expression with n components is *n-component confused*. In its current state, Camp's theory applies only to 2-component confused expressions because it uses 2-component logic. Keeping in mind the example of Fred and the ant farm, I introduce some terminology. Let a *query value* be one of the ways an expert (e.g., Sal or Sam) can answer when asked to evaluate a sentence. Thus, the query values are: Y, N, and ?. Let a *response value* be the group of query values assigned to a particular confused sentence. In the 2-component case, the response values are: YY, YN, Y?, NN, N?, and ??. Let a *semantic value* be one of the values assigned to each sentence by the logic. The response values are grouped into semantic values. In the 2-component case, the semantic values are: Y, N, YN, and ?.

If there were three big ants (Ant A, Ant B, and Ant C) in the ant farm instead of two, then 'Charlie' would be a 3-component confused expression. We would still use the same query values, but to evaluate Fred's confused arguments, we would need three experts and each confused sentence would be assigned a response value consisting of three query values. For example, if Fred utters 'Charlie is tired' and Ant A is tired, Ant B is tired, and Ant C is not tired, we ask the expert on Ant A whether Ant A is tired, we ask the expert on Ant B whether Ant B is tired, and we ask the expert on Ant C whether Ant C is tired. In this case, the response value for 'Charlie is tired' would be YYN. Once we have response values for Fred's sentences, we need a way of grouping them into semantic values and a way of evaluating his arguments for validity based on the semantic values of the sentences they contain.

Here are my suggestions. A 3-component logic has the following ten response values: Y??, YY?, YYY, N??, NN?, NNN, YYN, YNN, YN?, ???. These should be grouped in the same way as the response values for 2-component logic. That is, Y??, YY?, YYY are all Y; N??, NN?, NNN are all N; ??? is ?, YN? is YN; YNN is YNN; and YYN is YYN. That makes six semantic values: Y, N, YN, YYN, YNN, and ?. Thus, a 3-component confused expression requires a 3-component logic with six semantic values. A 4-component confused expression generates 15 response values and requires a 4-component logic with nine semantic values. In general, an n-component confused expression generates (n + 1)(n + 2)/2 response values and requires a logic with 3 + n(n – 1)/2 semantic values. However, some of these semantic values are redundant. For example, in 4-component logic, the value YYNN is equivalent to YN because these logics deal only with the ratio of Ys to Ns. Likewise, in a 6-component logic, YYYYNN is equivalent to YYN and YYNNNN is equivalent to YNN.

Recall the diagram for the semantic values of 2-component logic:

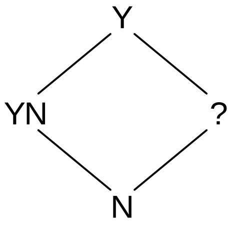


Figure 6.1 (2-component logic)

The following are the diagrams of the semantic values for 3-component logic and 4-component logic.

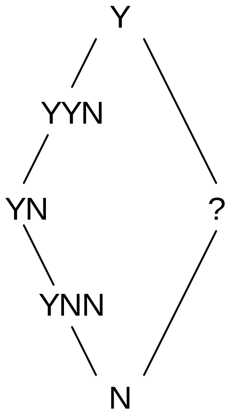


Figure 6.2 (3-component logic)

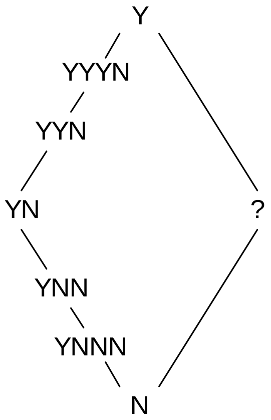


Figure 6.3 (4-component logic)

In these diagrams, implication goes from bottom to top.[6] If a line connects two semantic values, then a sentence with the semantic value nearer the bottom of the diagram implies a sentence with the semantic value nearer the top of the diagram. Thus, if every assignment of values to an argument with one premise assigns a value to the premise that is less than or equal to the value of the conclusion, then the argument is valid.

The justification for positing these relations between the semantic values of the 3-component logic and 4-component logic are as follows. In the case of 2-component logic, the four epistemic values are placed in the following groups: Y and YN are at-least-Y, N and YN are at-least-N. Valid arguments preserve at-least-Y and the absence of at-least-N (i.e., if an argument is valid and its premises are all at-least-Y, then its conclusion must be at-least-Y; if an argument is valid and its conclusion is at-least-N, then one of its premises must be at-least-N). In the 3-component case we have six semantic values: ?, Y, YYN, YN, YNN, and N. The relations between Y, N, YN, and ? should remain the same no matter how many extra semantic values are added. The problem is how to interpret arguments whose sentences are YYN or YNN. Grouping them into at-least-Y and at-least-N will not work because YN and YYN are both at-least Y and at-least N so if every assignment of values to a one-premise argument assigns YYN to the premise and YN to the conclusion then the argument would count as valid (according to validity for 2-component logic). But intuitively such an argument is invalid because YYN is “above” YN. According to our interpretation, a sentence assigned YYN has been endorsed by two of three experts and rejected by one, while a sentence assigned YN has been endorsed by one of three experts and rejected by one (the third is silent). Thus, an assignment of YYN to a sentence indicates greater profitability than an assignment of YN.

[6] Note that we cannot think of these lattices as approximation lattices in the way that Belnap suggests for the lattice for 2-component logic is an approximation lattice; see Belnap (1976b, 509-10) for more on this issue.

Similar considerations show that in an n-component logic, when two semantic values are combinations of ‘Y’s and ‘N’s, the one with a higher ratio of ‘Y’s to ‘N’s is above the other. Furthermore, if a semantic value has a combination of ‘Y’s and ‘N’s then it is incomparable to ?.

In an n-component logic, we want a valid argument to preserve the “height” of the semantic values. In other words, an argument is valid if and only if for every assignment of values, the premises are assigned semantic values that are “lower” than or equal to the semantic value assigned to the conclusion. To make this definition a bit more rigorous, I say that in an n-component logic, the *pure* semantic values are Y, N, and ?, while all the other semantic values are *mixed*. A mixed semantic value has a *Y-value* equal to the number of ‘Y’s in the name of the semantic value divided by the total number of letters in the name of the semantic value. The relations between the pure semantic values are: N<Y, ?<Y, N<?. Among the mixed values, if α and β are mixed values and α has a Y-value that is greater than or equal to the Y-value of β, then $\beta \leq \alpha$. Finally, if α is a mixed value then $N<\alpha$ and $\alpha< Y$. If υ is a map from a set of sentences into the set of semantic values for an n-component logic, then an argument composed of sentences from that set is *valid* if and only if for every valuation the semantic values assigned to the premises are less than or equal to the value assigned to the conclusion.

We can define negation, conjunction, and disjunction for n-valued logics as well. For negation we use the same definition as in 2-component logic but add that if $\langle p \rangle$ has a mixed semantic value with a Y-value of k, then $\langle \sim p \rangle$ has a mixed semantic value with a Y-value of 1-k. For example, if $\langle p \rangle$ is YYN, then $\langle \sim p \rangle$ is YNN; if $\langle p \rangle$ is YYYNN, then $\langle \sim p \rangle$ is YYNNN. We allow conjunction and disjunction to work the same way they do in 2-component logic. That is:

$\upsilon(\langle p \wedge q \rangle)$ = greatest lower bound of $\upsilon(\langle p \rangle)$ and $\upsilon(\langle q \rangle)$.

$\upsilon(\langle p \vee q \rangle)$ = least upper bound of $\upsilon(\langle p \rangle)$ and $\upsilon(\langle q \rangle)$.

In less rigorous terms, to determine the semantic value of a conjunction in an n-component logic, follow the lines downward in the diagram of the semantic values for that logic; where they *meet* is the semantic value of the conjunction. For example, if $\langle p \rangle$ is YYN and $\langle q \rangle$ is ?, then $\langle p \wedge q \rangle$ is N. To determine the semantic value of a disjunction in an n-component logic, follow the lines upward in the diagram for the semantic values of that logic; where they *join* is the semantic value of the disjunction. For example, if $\langle p \rangle$ is YNN and $\langle q \rangle$ is ?, then $\langle p \vee q \rangle$ is Y. The following are truth tables for negation, conjunction, and disjunction in 3-component logic:

p	~ p
?	?
Y	N
YYN	YNN
YN	YN
YNN	YYN
N	Y

∧	?	Y	YYN	YN	YNN	N
?	?	?	N	N	N	N
Y	?	Y	YYN	YN	YNN	N
YYN	N	YYN	YYN	YN	YNN	N
YN	N	YN	YN	YN	YNN	N
YNN	N	YNN	YNN	YNN	YNN	N
N	N	N	N	N	N	N

∨	?	Y	YYN	YN	YNN	N
?	?	Y	Y	Y	Y	?
Y	Y	Y	Y	Y	Y	Y
YYN	Y	Y	YYN	YYN	YYN	YYN
YN	Y	Y	YYN	YN	YN	YN
YNN	Y	Y	YYN	YN	YNN	YNN
N	?	Y	YYN	YN	YNN	N

The truth tables for $n > 3$ are straightforward to construct as well.

An important issue is: what is the relation between n-component logics and relevance logics? To show that 2-component logic is equivalent to R_{fde}, we made use of a result of Dunn's. Define $\langle p \rightarrow q \rangle$ to be *valid* if and only if for every valuation υ in a de Morgan lattice, $\upsilon(\langle p \rangle) \leq \upsilon(\langle q \rangle)$. Dunn proved that $\langle p \rightarrow q \rangle$ is valid if and only if $\langle p \rightarrow q \rangle$ is a theorem of R_{fde}. The semantic values of an n-component logic constitute a de Morgan lattice.[7] Thus, Dunn's theorem shows that the first-degree entailments of an n-component logic are theorems of R_{fde}.

We can extend an n-component logic to a full sentential logic or to a first-order predicate logic in the same way that we extended 2-component logic in section one. I will not go through the details. Instead, I want to move on to a much more difficult problem: how to deal with partial components.[8]

6.4 Partial Components

Recall that a partial concept is one whose range of applicability does not exhaust the domain in question. For example, it seems that numbers are not in the range of applicability of 'green'; hence, 'green' neither applies nor disapplies to numbers. We can say that applying or disapplying a partial concept to an object outside its range of applicability is a *category mistake*. Because many concepts are partial, an adequate theory of confused concepts needs to be applicable to confused concepts whose components are partial.[9]

[7] See Dunn and Restall (2001: 50-55) for a discussion of lattices in the context of relevance logic.

[8] There are other ways to extend the logic to n-components. I offer no assurance that the one presented in the text is the most intuitive. Another way of constructing n-component logics is to say that if ⟨p⟩ is ? and ⟨q⟩ has a mixed semantic value whose Y-value is greater than .5 then ⟨p⟩ implies ⟨q⟩, and if ⟨p⟩ has a mixed semantic value whose Y-value is less than .5 and ⟨q⟩ is ?, then ⟨q⟩ implies ⟨p⟩. One might be motivated to set up the logic in this alternative way if one believes that a sentence assigned a mixed semantic value whose Y-value is >.5 is more profitable than a sentence that is ? and that a sentence that is assigned a mixed semantic value whose Y-value is <.5 is less profitable than a sentence that is ?. The idea is that if more experts tell you "Yes" than tell you "No", then you are better off than if you were told nothing; if more experts tell you "No" than tell you "Yes" then you are worse off than if you were told nothing. I personally would rather be told nothing than be given conflicting advice by people I trust equally. It seems to me that this is a matter of personal preference.

[9] Camp suggests that one can explain category mistakes in terms of confusion, but he does not provide any details for how this can be done. If it could be done without an account of confused concepts with partial components, then one could use the theory of confusion at two different levels to explain confused concepts with partial components. However, I see little hope for this explanatory strategy. Instead, I provide a logic for confused concepts with partial components directly. It seems to me that this theory might be useful for explaining category mistakes, but I do not pursue that project here. See Camp (2002: ch. 19).

To have an example of conceptual confusion, consider Gil, who uses the term 'jade'. Unbeknownst to Gil, there are two "kinds" of jade: jadeite and nephrite.[10] Assume that Gil lives in a community where no one knows that 'jade' is confused. To justify the claim that 'jade' as it is used by Gil is actually confused, assume that Gil treats claims p and q as equally important in his use of 'jade', where the result of substituting 'nephrite' for 'jade' in p is true, but the result of substituting 'jadeite' for 'jade' in p is false, and vice versa for q.

If we assume that 'jadeite' and 'nephrite' are partial concepts whose ranges of applicability are the set of physical objects, then it makes sense to think of 'jade' as a partial concept with the same range of applicability. To simplify the discussion, I say that two concepts with identical ranges of applicability are *congruent*. It seems to me that if any two component concepts of a confused concept are congruent, then the confused concept and any one of its component concepts are congruent. I call such a confused concept *application-confused*. We still have no way of applying the theory of confusion to application-confused concepts because we have no query value for cases where the components neither apply nor disapply to an object, but it seems that one can adopt one's favored account of partial concepts and use it in conjunction with the theory of confused concepts. (The most intuitive account is that the sentences in question are truth-value gaps.[11]) However, if we assume that the range of applicability for 'jadeite' is the set of physical objects in China, while the range of applicability for 'nephrite' is the set of physical objects, then we run into trouble. I call a confused concept whose component concepts have different ranges of applicability *range-confused*. (Notice that the distinction between application-confused concepts and range-confused concepts is very

[10] Laporte (2004) casts doubt on the historical accuracy of this example.

[11] The weak Kleene scheme works well for accounts like this.

similar to the distinction between application-inconsistent concepts and range-inconsistent concepts—that is not a coincidence.)

Here is the problem range-confused concepts pose for my theory of confused concepts. Assume that ‘jade’ is range-confused in the way described above and that Gil presents an argument one of whose premises is ‘the statue of liberty is jade’. Because ‘jade’ is a 2-component confused concept, we use a 2-component logic with four semantic values to evaluate Gil’s confused arguments that contain occurrences of ‘jade’. Assume that our expert on jadeite is Jim and our expert on nephrite is Nancy. To assign a response value to the sentences of Gil’s argument, we ask Jim to evaluate ‘the statue of liberty is jadeite’. As I have defined ‘jadeite’, the statue of liberty is outside its range of applicability. As long as Jim is limited to Y, N, and ?, he have a difficult time providing a query value. It is inappropriate for him to evaluate the sentence in question in any of these ways. I suggest that we use a new query value: G. An expert replies to a query with G if and only if the object in question is outside the range of applicability of the concept. With this new option, Jim replies to our query with G. We ask Nancy whether the statue of liberty is nephrite and she replies with N. Thus, the response value for ‘the statue of liberty is jade’ is NG.

How should we evaluate Gil’s argument? One suggestion is to treat any confused sentence whose response value name includes a ‘G’ as a category mistake. On this view, a range-confused concept has a range of application that is the intersection of the ranges of application of its component concepts. Accordingly, for any range-confused concept, one can construct an application-confused concept that has the same application set, disapplication set, and range of applicability by defining component concepts that are analogous to the components of the original concept, but whose ranges of applicability are all the same. Thus, this approach to range-confused concepts assimilates range-confusion to application-confusion. However, these two types of confusion are distinct and should be treated as such. We need a way of evaluating arguments like Gil's that involve range-confused concepts. I call an n-component logic that allows G as a query value a *partial n-component logic*.

I propose to construct a logic for a 2-component range-confused concept and then generalize my findings to the family of partial n-component logics. For a 2-component range-confused concept, the response values will be: YY, YN, Y?, YG, NN, N?, NG, ??, G?, and GG. It makes sense to group them into the following semantic values: Y (YY and Y?), N (NN and N?), ? (??), G (GG and G?), YN, YG, and NG. Group the semantic values into the regular values—Y, N, ?, and YN—and the irregular values—G, YG, and NG. The most important principle for constructing a partial n-component logic is that it should preserve the implications present in the corresponding n-component logic. A second principle for deciding on the relations between these values is that if ⟨p⟩ is regular-valued and ⟨q⟩ is irregular-valued, then ⟨p⟩ does not imply ⟨q⟩. The justification is that we never want to introduce category mistakes into our reasoning—even if they are in the component concepts of our confused concepts. For certain values, we can allow ⟨p⟩ to imply ⟨q⟩ when p is irregular-valued and q is regular-valued. I assume that there are many ways of defining implications between irregular-valued sentences and regular-valued sentences, but the one I adopt is: if we replace the 'G' in the name of the semantic value with a 'Y', then a sentence with the original semantic value (before replacement) implies a sentence with the resulting semantic value (after replacement).

To make these claims more rigorous define the *Y-value* of a sentence as above, the *N-value* of a sentence as the number of 'N's in the name of its semantic value divided by the number of letters in the name of its semantic value, and the *G-value* of a sentence as the number of ‘G’s in the name of its semantic value (notice that the G-value is not analogous to either of the other two). A sentence with semantic value ? has no Y-value, N-value, or G-value. The first principle is that if the G-values of ⟨p⟩ and ⟨q⟩ are zero, then they behave just as they do on the corresponding n-component logic. The second principle is that if the G-value of ⟨p⟩ is less than the G-value of ⟨q⟩ then ⟨p⟩ does not imply ⟨q⟩. The third principle is that for ⟨p⟩ and ⟨q⟩ that have semantic values other than G, if the G-value of ⟨p⟩ is greater than or equal to the G-value of ⟨q⟩ and the Y-value of ⟨p⟩ is less than or equal to the Y-value of ⟨q⟩ and the N-value of ⟨p⟩ is greater than or equal to the N-value of ⟨q⟩, then ⟨p⟩ implies ⟨q⟩; otherwise, ⟨p⟩ does not imply ⟨q⟩. Using these rules, we can construct a diagram of the semantic values for a partial n-component logic. The following is the diagram for partial 2-component logic:

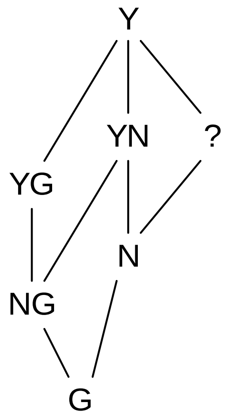


Figure 6.4 (Partial 2-Component Logic)

We can define the validity of an argument in partial 2-component logic in the familiar way: an argument is valid if and only if for every assignment of semantic values, the values of its premises are less than or equal to the value of its conclusion.

We define negation, conjunction, and disjunction as before:

p	~ p
?	?
Y	N
YN	YN
N	Y
YG	NG
NG	YG
G	G

∧	?	Y	YN	N	YG	NG	G
?	?	?	N	N	G	G	G
Y	?	Y	YN	N	YG	NG	G
YN	N	YN	YN	N	NG	NG	G
N	N	N	N	N	G	G	G
YG	G	YG	YN	G	YG	NG	G
NG	G	NG	NG	G	NG	NG	G
G	G	G	G	G	G	G	G

∨	?	Y	YN	N	YG	NG	G
?	?	Y	Y	?	Y	Y	G
Y	Y	Y	Y	Y	Y	Y	G
YN	Y	Y	YN	YN	Y	YN	G
N	?	Y	YN	N	Y	NG	G
YG	Y	Y	Y	Y	YG	YG	G
NG	Y	Y	YN	YN	YG	NG	G
G	G	G	G	G	G	G	G

These tables for conjunction and disjunction are defined in the same way that all the others have been: υ(⟨p ∧ q⟩) = greatest lower bound of υ(⟨p⟩) and υ(⟨q⟩); and υ(⟨p ∨ q⟩) = least upper bound of υ(⟨p⟩) and υ(⟨q⟩). The only exception is when one of the components of a disjunction is G, the entire disjunction is G. Thus, I have deviated somewhat from the standard account because of my interpretation of G. Negation is determined separately for each class of sentences with the same G-value: ⟨p⟩ has G-value k and Y-value m if and only if ⟨~ p⟩ has G-value k and Y-value 1-m.

Unfortunately, the semantic values of the partial n-component logics do not constitute de Morgan lattices; thus, Dunn's theorem cannot be used to demonstrate their relation to relevance logics. Of course, we can show that they have many of the properties of a relevance logic. For example, disjunctive syllogism is invalid and both the following inference rules are invalid:

p ∧ ~p ├ q

$q \vdash p \vee \sim p.$

We can introduce a conditional into a partial n-component logic in the same way that we introduced a conditional into n-component logics. A partial n-component logic can also be extended to a first-order predicate logic with universal and existential quantification defined as generalized conjunction and disjunction.[12]

The following are the diagrams of partial 3-component logic and partial 4-component logic:

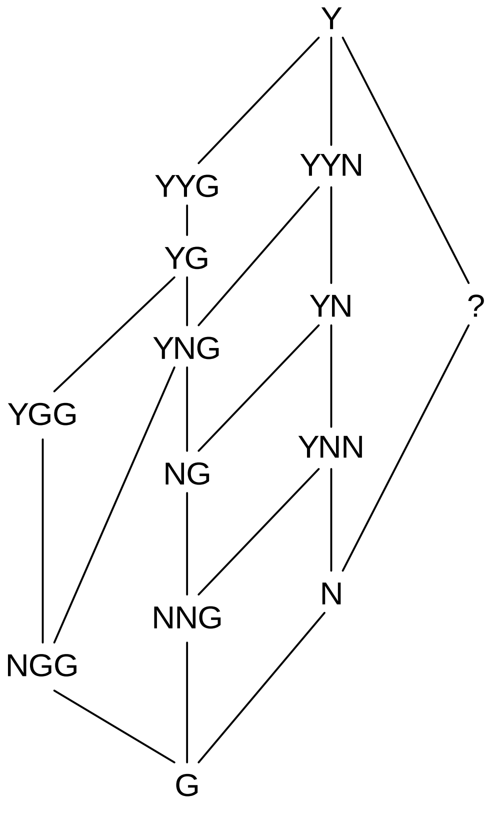


Figure 6.5 (Partial 3-Component Logic

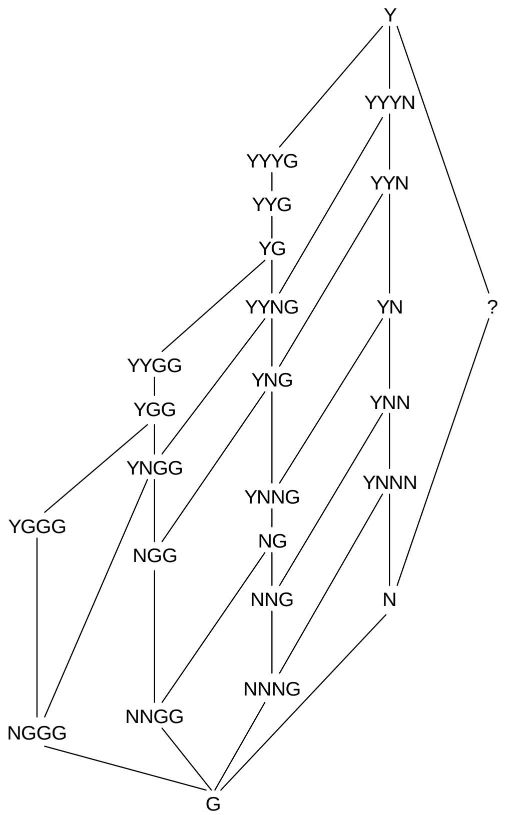


Figure 6.6 (Partial 4-Component Logic)

As n increases, the number of irregular values increases faster than the number of regular values.

[12] For the predicate logic we would allow the union of the extension and antiextension of an n-place predicate to be a proper subset of D^n.

There are no doubt other ways of constructing partial n-component logics. According to my account, the relation between an n-valued logic and a partial n-valued logic is similar to the relation between classical logic and a 3-valued logic with a weak Kleene scheme. According to the weak Kleene scheme, if a compound sentence has a component with value G, then the compound has value G. There is a 3-valued scheme called the strong Kleene scheme that does not have this feature. One could construct a family of partial n-component logics that are analogous to the strong Kleene 3-valued logic. Below are the diagrams for the "strong" versions of partial 2-valued logic and partial 3-valued logic.

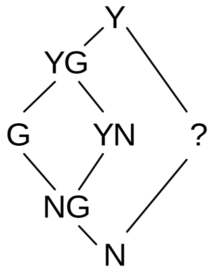


Figure 6.7 ("Strong" Partial 2-Component Logic)

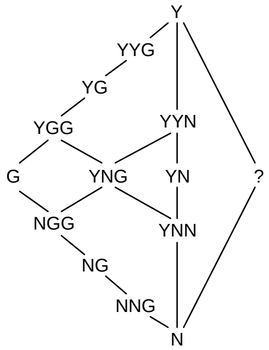


Figure 6.8 ("Strong" Partial 3-Component Logic)

6.5 A SEMANTIC THEORY FOR CONFUSED CONCEPTS

As we have seen, Camp presents a logic for confused expressions, but not a semantic theory for confused expressions.[13] In other words, Camp tells us which arguments that contain confused expressions are valid, but he does not tell us the meanings of the sentences that contain confused

[13] See Camp (2002: 220n4).

expressions. That is not meant to be a criticism of Camp. For my purposes, I need a semantic theory for confused expressions because I am trying to show that if one treats truth as an inconsistent concept, then one can have an internalizable semantic theory for truth. My view is that inconsistent concepts are confused. Thus, in order to construct a semantic theory for truth, I need a semantic theory for confused expressions.

In order to construct a semantic theory for confused expressions, I make use of Brandom's theory of meaning, which is an inferential role theory of meaning; it explains the meaning of a sentence in terms of its role in inferences and it explains the meanings of subsentential expressions in terms of their contribution to the inferential roles of the sentences in which they occur. I have two reasons for choosing an inferential role theory as the basis for a semantic theory for confused expressions: (i) confused sentences have no truth-values, which makes it difficult to use a truth conditional theory of meaning, and (ii) I have already endorsed a particular logic for confused expressions, which can be used to determine the inferential role of confused sentences.

6.5.1 BRANDOM'S THEORY OF MEANING

The heart of Brandom's *Making It Explicit* is a theory of discursive practice. We are all familiar with discursive practices, for they are the practices in which participants behave in a way that is sufficient to confer content on some of their performances, mental states, and products. Although 'content' could use a sharp definition, I am not going to provide one. Suffice it to say that 'content' is used in a way that is similar to the way 'meaning' is used, except that where 'meaning' applies to linguistic entities alone, 'content' applies to mental ones as well (e.g. mental states, attitudes, etc.).

Brandom divides his theory of content into two parts: a theory of semantic phenomena and a theory of pragmatic phenomena. One of his fundamental commitments is that one should explain the former in terms of the latter, which is a descendant of the view that meaning should be explained in terms of use. Thus, semantic phenomena (content, truth, reference, validity) are explained in terms of the way the things that bear content are used.[14]

I address Brandom's semantic theories and pragmatic theories in turn. As I mentioned, his semantic theory belongs to a family called *conceptual role semantics*. Members of this family explain meaning or content in terms of the conceptual role of the thing that bears the meaning or content. Brandom's version takes the conceptual role of a content-bearer to be its role in inference. He takes the primary content-bearers to be sentence tokens and the primary notion of inference to be material inference, which is a relation between two content-bearers that holds (in part) because of the content they bear (as opposed to formal inference which holds because of the form of the content-bearers). Thus, the content of a sentence token is its inferential role.

Brandom distinguishes between three types of inferential relations: commissive, permissive, and incompatibility (I explain these terms when I turn to his pragmatic theory). Accordingly, the inferential role of a sentence has three parts—one for each of the inferential relations in which it participates. We can think of the incompatibility role as a set of ordered pairs of sentences that are incompatible. The commissive and permissive parts can each be thought of as sets of inferential antecedents and inferential consequents. The *antecedents* of a sentence are the sentences from which one can infer the sentence in question (let us call it *p*) and the *consequents* are the sentences that one can infer from it. The antecedents of *p* form a set whose members are sets of sentences from which *p* follows. The consequents of *p* form a set of

[14] See Brandom (1994: chs. 1-2).

ordered n-tuples. The first member of each n-tuple is a sentence that follows from *p* and the other members of each n-tuple are the premises besides *p* needed to derive the first member. Thus, the inferential role of a sentence p, will be: {p's commissive antecedents, p's commissive consequents, p's permissive antecedents, p's permissive consequents, p's incompatibilities}. The inferential role of a sentence depends on both which sentences are available to serve as auxiliary premises and which inferences are correct. Brandom assumes that each member of a discursive practice takes everyone else to agree on the latter.[15] (I reject this assumption later in this chapter.)

Brandom's pragmatic theory takes as primitives the notions of *deontic status* and *deontic attitude*. Statuses come in two flavors: *commitments* and *entitlements*. The former are similar to obligations and the latter are similar to permissions. There are three types of attitudes: *attributing*, *undertaking*, and *acknowledging*. One may attribute, undertake, and acknowledge various commitments and entitlements.

There are several different kinds of commitments that correspond to aspects of discursive practice. *Doxastic commitments* correspond to assertions and beliefs, *inferential commitments* correspond to reasons, and *practical commitments* correspond to actions. The members of a discursive practice keep track of each other's commitments and entitlements. Brandom adopts Lewis's explanation of this behavior in terms of scorekeeping.[16] At a given moment in a conversation, the score is just the commitments and entitlements associated with each participant. Each member of the conversation keeps score on all the participants (including herself). Every time one of the participants undertakes (implicitly adopts), acknowledges

[15] Brandom (1994: ch.2).

[16] See Lewis (1979).

(explicitly adopts), or attributes (takes another as if he adopts) a commitment or entitlement, it changes the score. I will refer to these as *scorekeeping actions*.

Brandom bases his pragmatic theory on the idea that the use of a linguistic item is the way it changes the score of a conversation. Because Brandom emphasizes the normative dimension of content, he defines the *pragmatic significance* of a sentence as the way it *should* affect the score of a conversation in which someone utters it. Pragmatic significance has two aspects—the circumstances of application and the consequences of application. The former consists of the scores of conversations in which it is legitimate to utter the sentence in question. The latter is the scores that should result from a legitimate utterance of it.[17]

There are two important senses in which Brandom's semantics answers to his pragmatics. First, the inferences that constitute the content of a sentence are explained in terms of commitments and entitlements. A *commissive inference* is one for which if one is committed to its premises, then one should be committed to its conclusion as well. If one is entitled to the premises of a *permissive inference*, then one should be entitled to its conclusion too. Two sentences are *incompatible* if commitment to one precludes entitlement to the other. The participants of an inferential practice acknowledge inferential commitments by using some sentences as reasons for others. Second, given the force of an utterance, the content of the sentence uttered determines its pragmatic significance. That is, once the members of a discursive practice determine that a given utterance has a certain force, they can use the content of the sentence uttered (its inferential role) to determine how it should change the score of the conversation (its pragmatic significance).

[17] Brandom (1994: 180-198). Recall that these notions figure in my argument that confused sentences have no truth-values in section 5.2.3 of Chapter Five.

For Brandom, the paradigmatic use of a sentence is an assertion. Consequently, his model of discursive practice is one in which the members make various assertions. He assumes that this model can be extended to include all the other types of speech acts. When a person makes an assertion, she sets off a chain reaction of scorekeeping actions by each member of the conversation. Three important features of assertions govern these scorekeeping actions. First, when someone makes an assertion, she acknowledges a doxastic commitment. She also undertakes all the commitments and entitlements that follow from the one acknowledged. Second, a successful assertion (i.e., one in which the asserter is entitled to the commitment acknowledged) entitles other members of the conversation to undertake the same commitment. Successful assertions present commitments for public consumption. Third, the asserter takes responsibility to justify the assertion by giving reasons for it should the need arise. In general, assertion displays a *default and challenge structure* in which many assertions carry default entitlement that another member of the conversation can challenge.[18]

Brandom extends his basic model in several different ways. He accounts for the commitments undertaken and acknowledged in perception by explaining perceptual reliability in terms of a special kind of inference. To account for action, Brandom introduces *practical commitments*, which are involved in inferences and can have entitlement associated with them. He treats actions as acknowledgments of practical commitments and presents a rudimentary action theory in terms of this idea. He uses a notion of *substitution* to extend his account of inferential role from sentences to subsentential expressions and a notion of *recurrence* to extend it from subsentential expressions to context-sensitive performances. He also introduces

[18] Brandom (1994: 167-179).

scorekeeping actions to account for all of these subsentential semantic phenomena.[19] In this chapter, I can deal only with the sentential level.

6.5.2 A Pragmatic Theory for Confusion

The point of this subsection is to present an extension of Brandom's theory of content. At the semantic level, the extension allows members of a discursive practice to disagree about which inferences are correct. It will also allow them to adopt different semantic stances (i.e., use different standards when evaluating inferences). At the pragmatic level, the extension allows scorekeepers to acknowledge, undertake, and attribute inferential commitments to one another. Although Brandom's model already includes inferential commitments, he assumes (to simplify the theory) that each member of a discursive practice attributes the same ones to everyone else. The extension also introduces a new type of status: scorekeeping commitments. These allow scorekeepers to change the way they keep score on one another.

These additions allow Brandom's model of content to explain what it is to adopt a semantic stance. When a person adopts a semantic stance, one commits oneself to an inferential standard for use in assessing someone's inferential behavior. I explain adopting semantic stances in terms of acknowledging scorekeeping commitments. The reason for this strategy is, of course, to comply with his principle that pragmatic phenomena should explain semantic ones. Once complete, the extension of Brandom's theory of content will yield a pragmatic version of Camp's theory of confusion and it will provide a semantic theory for confused concepts.

I present the extension of Brandom's theory of content in two parts: the account of inferential commitments and the account of scorekeeping commitments. They are combined to

[19] See Brandom (1994), chapters 4, 6, and 7, respectively.

explain what it is to adopt a semantic stance in general and the semantic stance appropriate for the confused in particular. The following are two reasons his theory needs the extension.

First, people disagree on which inferences are correct. Brandom explains this disagreement in terms of differences in beliefs. According to Brandom, people disagree about which sentences follow from a given sentence not because they accept different inferences but because they accept different potential premises. One's views on what follows from some claim will depend on both the inferences one endorses and the sentences one has available to use as premises.[20] However, people also disagree about which inference rules are correct. One cannot explain this disagreement in terms of differences in beliefs. If Brandom's theory of content is to describe actual discursive practices then it will have to allow practitioners to endorse different inferential standards.

Second, discursive practitioners adopt semantic stances with respect to one another. We do not hold each other to the same inferential standards. The standard one uses for assessing inferences varies from person to person and context to context. If Brandom's account of discursive practice is to be realistic, it must capture this important aspect of our inferential behavior. There is a difference between treating someone as if he is inferring incorrectly according to his own standards and treating him as if he has adopted the wrong standards. One needs to look no further than common philosophical debates for evidence of this phenomenon. For example, it is *appropriate* for a classical logician to treat an intuitionist as if he has made a simple logical error if the intuitionist's argument employs double negation, even though the classical logician accepts this inference rule. On the other hand, it is *inappropriate* for an intuitionist to treat a classical logician as if she has made a simple logical error if the classical logician's argument employs double negation, even though the intuitionist rejects this rule. The

[20] Brandom (1994: 357).

debate between intuitionists and classical logicians that we find in the philosophical literature is one in which each finds faults with the other's inferential behavior. However, they take one another to have adopted the wrong inferential standards. An account of adopting a semantic stance allows Brandom's model of discursive practice to explain this phenomenon.

The two reasons given above are related. I argued that if Brandom's theory is to account for the fact that humans endorse different inferences, then it has to allow scorekeepers to acknowledge, undertake, and attribute inferential commitments to one another. Further, not only do people endorse different inferences, but we also evaluate others according to different standards of what counts as a good inference. The two phenomena go hand in hand. If I can attribute inferential commitments to you that are different from those I acknowledge, then I need a way of judging whether you have followed your own inferential commitments. Semantic stances fit the bill. By adopting a semantic stance on you, I assess your arguments according to inferential standards that I might not accept. Thus, allowing scorekeepers to disagree about inferential correctness and allowing them to adopt semantic stances go hand in hand. A discursive practice in which scorekeepers acknowledge, undertake, and attribute inferential commitments is one in which scorekeepers adopt semantic stances.

6.5.2.1 Inferential Commitments

My goal in this subsection is to extend Brandom's scorekeeping pragmatics to conversations in which participants endorse different inferences. The way to accomplish this is to permit scorekeepers to keep track of each other's inferential commitments. An *inferential commitment* is a type of deontic status that one can undertake, acknowledge, or attribute; it is just like a doxastic commitment or a practical commitment in this respect. One can be entitled to inferential commitments as well. There are, of course, differences between inferential commitments and doxastic commitments. One expresses a doxastic commitment by uttering an assertion, while one expresses an inferential commitment by treating one doxastic commitment as a good reason for another. It might seem that one could express an inferential commitment by asserting that one sentence follows from another. Although I do not want to rule this out, I do not want the possibility of expressing inferential commitments to depend on the presence of logical vocabulary. I want a model of scorekeeping that incorporates differences of opinion about inferential commitments from the start.

I need to address a number of other issues surrounding inferential commitments. First, do they participate in inferential relations? That is, can one infer one inferential commitment from another? It seems to me that the answer is *yes*. For example, an inferential commitment expressed by <<something is flat ∴something is flat>>[21] follows from the inferential commitment expressed by <<something is flat and brown ∴something is flat>>. This issue is important for formulating the norms that govern scorekeeping practice. For example, Apu might want to say that if Manjula is a reliable observer of red things, then he is too. Recall that inferential commitments explain the status of observational reliability. Thus, Apu's formulation of the scorekeeping norm expresses an inferential commitment that holds between two inferential commitments. The fact that inferential commitments participate in inferences implies that scorekeepers must keep track of the inferential commitments acknowledged and those undertaken by each member of a conversation.

Another issue is the way in which one can come to be entitled to an inferential commitment. We can extend the default and challenge structure to them in a straightforward

[21] The double angle convention is used to construct names for arguments.

way. When someone makes an assertion, is challenged on it, and makes another assertion that is intended to serve as a reason for the first, a member of the audience can challenge the asserter in two different ways. An audience member can make a *doxastic challenge* in which he challenges the asserter to demonstrate entitlement to the doxastic commitment expressed by his second assertion; or an audience member can make an *inferential challenge* in which he challenges the asserter to demonstrate entitlement to the inferential commitment expressed by his use of the second assertion as a reason for his first. There will be at least two types of inferential challenge. One would challenge the speaker's inferential standards, while the other is for the case where the speaker violated his own professed standards.

One might make a case for the claim that one can have default entitlement to an inferential commitment based on one's status as a reliable reporter. However, it seems doubtful that a member of a discursive practice that does not contain logical vocabulary will be able to provide a satisfactory response to an inferential challenge. Nevertheless, a scorekeeper in such a discursive practice can register the fact that he does not endorse the inferential commitment undertaken by the asserter. In a more advanced discursive practice, one can justify inferential commitments and inherit them by testimony. (Debates about intuitionism and relevantism provide a number of good examples of each of these discursive phenomena.)

One important consequence of this addition to Brandom's scorekeeping pragmatics is that propositional content will be doubly perspectival. Brandom is already committed to the view that people who acknowledge different doxastic commitments will disagree about the inferential role of a claim (i.e. its content). If one accepts the claim that scorekeepers can differ on which inferences they endorse as well, then propositional content will be relative to a set of doxastic commitments and to a set of inferential commitments.

6.5.2.2 Scorekeeping Commitments

I need to introduce a new type of commitment into Brandom's pragmatic theory to explain what a scorekeeper is doing when she adopts a semantic stance. A *scorekeeping commitment* is a type of practical commitment—a commitment to action. That is, one performs an action by acknowledging a practical commitment. By acknowledging a scorekeeping commitment, one performs a special type of action—one keeps score. Undertaking a scorekeeping commitment is a way of saying, "I am going to keep score in such and such a way." It is a commitment to future scorekeeping actions. One can, of course, change the way one keeps score. In this case, one acknowledges a new scorekeeping commitment.

For the most part, scorekeeping commitments obey the rules for practical commitments. Thus, one can acknowledge, undertake, and attribute scorekeeping commitments. They participate in inferences and are susceptible to entitlement as well. The fact that scorekeeping commitments participate in inferences implies that scorekeepers will have to keep track of the scorekeeping commitments acknowledged and those undertaken by each member of a conversation.

One can distinguish several different types of scorekeeping commitments. There are those that affect how one keeps score on oneself and those that affect how one keeps score on others. (Example of a change in the latter: "I'm going to pay more attention to Otto's attitudes toward Becky.") There are those that affect the way one inherits commitments and entitlements from others. (Example: "I'm going to be less gullible.") Some scorekeeping commitments pertain to the relation between different types of commitments. For example, one can acknowledge a scorekeeping commitment to treat only those who accept the claim that monkeys do not grow on trees as possessors of the concept of a monkey. That is, a scorekeeper might interpret a person's use of 'monkey' as meaning *monkey* only if the scorekeeper attributes to this person the doxastic commitment associated with the claim that monkeys do not grow on trees. Otherwise, the scorekeeper will treat the person's term 'monkey' as if it means something else (or nothing at all). One acknowledges one of these scorekeeping commitments when one calls a sentence "meaning-constitutive". Similar scorekeeping commitments pertain to attributions of analyticity, definition, etc. There are scorekeeping commitments that are appropriate only for the one who undertakes them and those that are appropriate for everyone in a particular situation. For example, if one member of a three-person conversation realizes that one of the other members is confused on some topic, and realizes that the third member recognizes the confusion as well, then the first will adopt a scorekeeping commitment with respect to how to assess the confused person's inferences. Moreover, the first treats this scorekeeping commitment as one the other (non-confused) member of the conversation ought to adopt as well. The semantic stance associated with confusion is one that is appropriate for anyone who deals with a confused person. This list is far from complete but I hope it helps flesh out the idea of a scorekeeping commitment.

An important issue is entitlement to scorekeeping commitments. As with all commitments, there should be a default and challenge structure associated with scorekeeping commitments. For example, Camp presents a reading of Locke according to which he is confused.[22] A participant in a conversation with Camp might say to him, "Locke does not confuse acts and objects, so stop treating him as if he does." Camp would then have an opportunity to justify his scorekeeping commitment. The way entitlements to scorekeeping commitments are passed from person to person will be a bit tricky. Since scorekeeping

[22] Camp (2002: 191-217).

commitments are practical commitments, it will depend on the role entitlement plays for practical commitments. I remarked at the end of the previous paragraph that some scorekeeping commitments will have inheritance structures such that if one member of a conversation entitles himself to one of these scorekeeping commitments, then the others are entitled to endorse it as well. I will have to leave the details for some other occasion.

6.5.2.3 Semantic Stances

Semantic stances involve standards by which one assesses arguments for validity. (I restrict my attention to deductive inferences.) I should mention that when someone treats another as confused, she adopts *one type* of semantic stance, and when someone adopts a semantic stance, he acknowledges *one type* of scorekeeping commitment. There are many other types of scorekeeping commitments and many other types of semantic stances.

When a member of a discursive practice adopts a semantic stance, she acknowledges a scorekeeping commitment. The content of her scorekeeping commitment is that she will evaluate the inferences of some other scorekeeper according to some standard. Obviously, scorekeepers always employ some set of inferential commitments to assess inferences. Thus, scorekeepers always employ some semantic stance or other. We can think of the most common one as a default stance. Most likely, the default stance will be one that takes everyone to endorse the same inferential commitments. The default stance corresponds to a scorekeeping commitment to assess others' inferences according to one's own inferential commitments. When a scorekeeper adopts a different semantic stance, she acknowledges a new scorekeeping commitment. She commits herself to evaluate the inferences of another according to some inferential standard that she might not endorse.

6.5.2.4 Pragmatics for Confusion

On Camp's account of confusion, someone who interacts with a confused person should adopt a semantic stance according to which she does not attribute truth-values to the confused sentences, and she assesses them according to whether they preserve profitability. The person adopting the new semantic stance uses one of the n-component logics to track profitability. At the pragmatic level, adopting this semantic stance corresponds to acknowledging a specific scorekeeping commitment.

In the interest of space, I do not present any of the substitution and recurrence structures that allow Brandom to extend his theory of content from the sentential level to the subsentential level. Thus, although confusion is essentially a subsentential phenomenon in that confusion pertains to subsentential expressions, I deal with confused sentences only.

Let us return to Fred, Ginger, and the ants. Assume that Ginger has decided that Fred is confused. Any sentence Fred utters containing 'Charlie', 'the big ant', etc. will count for Ginger as a confused sentence. Any argument that contains confused sentences is a confused argument. In semantic terms, once Ginger has decided that Fred is confused, she adopts a particular semantic stance toward him. I refer to it as the *confusion stance*. When Ginger adopts the confusion stance, she decides not to attribute truth-values to Fred's confused sentences. (Recall that no such assignment can be inferentially charitable.) Further, she assigns semantic values from the 2-component logic to Fred's confused sentences in an effort to assess his arguments for profitability preservation. To do so, she must either have the authority to play the roles of Sal and Sam or else have access to someone who does. Once Ginger assigns the semantic values, she can evaluate Fred's confused arguments for validity.

In pragmatic terms, once Ginger has decided that Fred is confused, she acknowledges a scorekeeping commitment. It is a commitment to keep score on Fred in a certain way. In order to demonstrate the content of this commitment, assume that Fred utters a sentence, *p*, as the conclusion of an argument whose only premise is *q*. Assume also that Ginger has decided that *p* and *q* are confused sentences. Ginger decides that *p* is an assertion. She understands its content and attributes to Fred a doxastic commitment that corresponds to it. Assume that Fred is not default entitled to it and he has not acquired it by testimony. Ginger must decide whether Fred's argument, <<q∴p>>, entitles him to *p*.

The scorekeeping commitment Ginger acknowledges has four aspects. First, she refuses to attribute truth-values to Fred's confused sentences. For Brandom's pragmatic theory, this amounts to a refusal to acknowledge either the doxastic commitments she attributes to Fred (even if he turns out to be entitled to them) or the doxastic commitments that correspond to their negations. Thus, she must disengage from an important part of what it is to treat an utterance as an assertion. Although Fred is making assertions, his commitments are not fit for public consumption. Second, Ginger treats Fred as if he has undertaken new inferential commitments. These inferential commitments correspond to those deemed valid by the 2-component logic. These inferential commitments will most likely be different from the ones Ginger acknowledges. Note that Fred would probably not acknowledge these inferential commitments either. However, by virtue of being confused, he has undertaken them (according to Ginger). Third, she uses these inferential commitments to assess Fred's confused arguments. To do so, she must acknowledge a doxastic commitment to the effect that she has access to an authority on the topic about which Fred is confused. She now consults this authority (which might just be her) and acknowledges doxastic commitments that correspond to the substitutional variants of Fred's confused sentences (the sentences that result from replacing ‘Charlie’ with ‘Ant A’ or ‘Ant B’). She uses these doxastic commitments to attribute epistemic values from the 2-component logic to Fred’s confused sentences. She then evaluates Fred’s argument (<<q∴p>>) according to the inferential commitments she attributed to him in the second stage. Fourth, she uses the results of the previous two stages to determine whether she should attribute entitlement to p. If she takes Fred to be entitled to q, and she takes <<q∴p>> to be valid by the logic in question, then she takes Fred to be entitled to p. Remember that she does not take this attribution of entitlement to authorize anyone else to acknowledge *p*. (One consequence of this account of the pragmatics of confusion will be that the notion of entitlement is split into a weak version that does not entitle others to adopt the same commitment and a strong version that does.)

It is essential to appreciate that the scorekeeping commitment Ginger acknowledges undermines an important aspect of assertion. Camp argues that when interpreting the confused, there is a tension between two aspects of understanding: assessing reasons and assessing beliefs. Brandom’s model of assertion fuses these two components of understanding. He emphasizes the fact that, in general, understanding someone’s belief requires not only deciding whether to adopt it, but also appreciating his reasons for it as well. For Brandom, if I think you have a good reason for your belief, then I have good reason to accept it too (other things being equal). In other words, Brandom builds inferential and doxastic charity into his model of assertion. However, in the confusion example, Ginger can think that Fred has a good reason for his confused belief only if she refuses to even consider whether she should accept it or reject it. Inferential and doxastic charity are incompatible in the presence of confusion. If Brandom’s model of assertion is correct, then inferential and doxastic charity must coincide in general. That is, one cannot attribute confusion to everyone and still be participating in a discursive practice.

Thus, scorekeeping commitments for confusion must be exceptions to the norm. Adopting the confused position is a discursively advanced thing to do.[23]

6.5.3 A Semantic Theory for Confusion

I have followed Camp's theory of confusion by endorsing a relevance logic as the standard by which one ought to interpret the claims of the confused and I have presented several alterations of Brandom's pragmatic theory that allow one to explain in scorekeeping terms what is done by an interpreter who takes another to be confused. This is a pragmatics for confusion attributions in the sense that it provides an account of how language users should treat utterances that they take to be confused. It is not an account of how language users should use confused concepts. I hope that no such account is necessary (aside from saying exactly how one goes about discarding a concept and replacing it with new ones). The logic Camp advocates for confusion is a logic for confused sentences in the sense that it specifies which confused arguments are valid. It is not a logic for confusion attributions in the sense of specifying which arguments that contain confusion attributions are valid. I have already said that I do not provide a pragmatics for confused expressions because I do not anticipate any use for such a thing. I can also say that I will not provide a logic for confusion attributions, not because I do not anticipate any use for one, but because they do not seem to call for special treatment. Likewise, I do not present a semantics for confusion attributions, mostly because I do not anticipate any big problem with them. The semantics I present in this section is a semantics for confused expressions.

[23] See Camp (2002: ch. 18).

6.5.3.1 BRANDOM'S SUBSTITUTION SEMANTICS

I find chapter six of Brandom's *Making It Explicit* to be the most difficult in the book. Unfortunately, my confusion semantics depends on some of the tools forged in Brandom's foundry. Thus, I first explain what goes on in the first section of chapter six. I do not have a good term for the theory that is presented there, so I have just called it *Brandom's Substitution Account*. It allows one to extract several different notions of content for the sentences of some bit of discourse from a logic for these sentences.

The first thing we need is a distinction between designated value and multivalue for multivalue logic. Most of us are familiar with classical logic that uses two truth values, true and false. These are multivalues. One can construct logics with lots of multivalues. I call them *truth-values* throughout this dissertation. Designated values are also familiar but not as distinct from multivalues. Truth is designated in classical logic; however, one can construct a multivalue logic in which more than one multivalue is designated. A multivalue is designated if it is prized in some sense. We want our assertions to be true and our inferences to preserve truth; thus, truth is designated. Brandom's use of multivalue logic requires a notion of *designated value*, instead of just designatedness. Even in logic with more than two multivalues, there are almost always only two designated values: designated and not designated. In classical logic, there are two multivalues, true and false, and two designated values, designated and not designated. Truth is designated, falsity is not designated. One can construct a logic with three multivalues, true, false, and gappy, in which true is designated and gappy and false are not designated (this is the standard way to do three valued logic). In Brandom's account there can be more than two designated values. This is a hard thing to grasp because we are used to thinking of multivalues as having their designatedness "built in." That is, we often think of multivalues as being designated or not designated. Nevertheless, the substitution account depends on a system with lots of multivalues and lots of designated values.

Formal logic is formal in one sense because the sentences involved display their semantic features syntactically. Many logics deal with sentences whose multivalues are determined by the multivalues of their components. Most also include a notion of validity according to which a sentence is valid if and only if it is designated no matter what the multivalues of its components. In classical logic, validity is logical truth. A sentence is a logical truth if and only if it is true no matter whether its components are true or false. Here, truth plays the role of a multivalue and a designated value. In the general case, one must distinguish between the multivalue and the designated value.

The substitution account is designed to apply to any linguistic item that has linguistic items as components. Thus, it applies to sentences that have other sentences as components, it applies to sentences that have words as components, and it applies to arguments that have sentences as components. The normal way to do things is bottom-up. That is, one specifies the multivalues of the components and then calculates the designated values of the compounds. The substitution account is top-down in the sense that one calculates the multivalues of the components from the designated values of the compounds. The essence of the account is that two components have the same multivalue if substituting one for the other does not change the designated value of the compound. We assume that designated values are attached to multivalues in the sense that if two items have the same multivalue then they have the same designated value as well.

The first task is to distinguish freestanding from ingredient content and use the former to derive the latter. Both pertain to sentences; *freestanding content* is the content that a sentence has when it is on its own, and *ingredient content* is the content of a sentence when it serves as a component in some larger compound. A sentence's ingredient content is the contribution it makes to the freestanding content of the compound in which it is a component. Let S be a set of compound sentences, s_1-s_n, and C be a set of atomic sentences, c_1-c_m, that are the components of the sentences in S; assume that all belong to the same language. If we are given the designated values for s_1-s_n, then we can derive their multivalues in the following way. For any two sentences c_i and c_j of C, they have the same multivalue if and only if for every sentence s_k of S, substituting c_i for c_j or c_j for c_i in s_k does not change the designated value of s_k. Let the designated values of the members of S be their freestanding contents and the multivalues of the members of C be their ingredient contents. Ingredient contents are equivalence classes of sentences on this account. Obviously, the ingredient contents will depend on the embedding contexts available.

Let us turn our attention toward arguments now. Think of an argument as a compound whose components are sentences. Let A be a set of arguments and S be a set of sentences that feature in those arguments. Assume that all the arguments in A are composed of sentences from S. If we are given the designated values of the arguments in A then we can extract the multivalues of the sentences in S using the substitution account. Any two sentences s_i and s_j have the same multivalue if and only if for every argument in A, substituting one for the other does not change its designated value. If we use two designated values (correct, incorrect) for the members of A then the multivalues are *freestanding inferential contents* of the sentences of S. We can then generate the *ingredient inferential contents* of the components of those sentences of S that are compounds by using the substitution account one more time. Just as above, let the compound sentences of S be S′ and the compounds that are contained in them be C. Let the designated values be the freestanding inferential contents of the members of S′ and the multivalues be the ingredient inferential contents of the members of C. Obviously, the freestanding inferential contents will depend on which argument contexts are found in the members of A and the ingredient inferential contents will depend on which embedding contexts are found in the members of S′.

The substitution account allows Brandom to construct two hierarchies. The assertional hierarchy has two levels and consists of sentences and their freestanding assertional content on the top level and sentences and their ingredient assertional content on the bottom level. The inferential hierarchy has three levels, with the arguments and their designated values on the top level, sentences with freestanding inferential contents on the middle level, and sentences with ingredient inferential contents on the bottom level. Note that the freestanding inferential content of a sentence need not be the same as its freestanding assertional content and its ingredient inferential content need not be the same as its ingredient assertional content. Brandom does claim that if two sentences have the same ingredient inferential content, then they will have the same ingredient assertional content.[24]

6.5.3.2 Semantics For Confusion

I want to use the substitution account to generate contents for confused sentences. Let us revisit Fred and Ginger. Assume that Ginger is just where we left her—she has decided that Fred is confused and she has identified the confused sentences of Fred’s. She has also adopted a semantic stance on Fred and she has performed a plethora of scorekeeping actions to keep track of Fred’s confused sentences in accordance with the inferential standard she has adopted for

[24] Brandom (1994: 351).

them. The task now is to provide meanings for Fred's sentences. Given the substitution account and the designated values of Fred's confused arguments, it is fairly straightforward. Ginger should assign Fred's confused sentences multivalues based on whether substitution changes the designated value of the argument in which they occur. Obviously, Ginger should construct every argument possible using some specified set of confused sentences, assign them designated values based on the 2-component logic, determine the multivalues of the confused sentences using the substitution account, and treat multivalues as freestanding inferential contents. If she so chooses, she can extract ingredient inferential content by using the substitution account once more.

Obviously, this account of the semantics for Fred's confused sentences will not generate freestanding assertional contents. It seems to me that Ginger could determine such things, but they will have to be derived from the confusion logic and the confusion semantics. Any confused sentence to which Fred is entitled according to Ginger's confusion logic will be assertible. Obviously, Fred and Ginger will differ on which sentences are assertible and on which arguments are correct. One can use the substitution account to generate whatever one wants in terms of content for him, content for her, assertibility for him, assertibility for her, etc. I do not want to come down on which ones of these will count as the best way for her to proceed. Rather, I want to present a set of tools that can be used in the face of confusion.

6.6 Confused Concepts and Inconsistent Concepts

Now that I have presented and extended the theory of confused concepts, I make six points about inconsistent concepts. First, conceptual inconsistency is explained in terms of conceptual confusion and conceptual confusion is explained in terms of conceptual confusion attributions.

Thus, conceptual inconsistency is explained in terms of what it is to treat something as conceptually inconsistent. Hence, a concept is inconsistent if and only if it is appropriate to adopt a certain semantic stance toward the employer of that concept. We now have an account of what it is to adopt a semantic stance, which is explained in terms of scorekeeping pragmatics.

Second, when one adopts the semantic stance appropriate for conceptual inconsistency, one uses a certain logic to evaluate the confused arguments in question and one keeps score on the confused in a certain way. Using these methods, one can attribute meanings to the confused sentences in question and determine whether assertions of confused sentences are warranted or unwarranted.[25]

Third, because I have extended the theory of confusion to handle n-component confusion and partial components, my theory of inconsistent concepts is applicable to both application-inconsistent concepts and range-inconsistent concepts and it is applicable to inconsistent concepts with more than two components. I have advocated a certain family of logics for confused expressions and these carry over to inconsistent expressions well. In the simplest case, an inconsistent concept will have two components that are both completely defined (i.e., they have empty ranges of inapplicability). In this case, a 2-component logic is the appropriate one. In cases where an inconsistent concept has n components (n>2), a logic with more semantic values will be appropriate (e.g., 3 components requires 6 semantic values, and 4 components requires 8 semantic values). In cases where an inconsistent concept has partial components, more complex logics are required. If all the components of an inconsistent concept are partial but have identical ranges of applicability, then one needs to add only one semantic value (G) to the logic for category mistakes. However, if an inconsistent concept is range-inconsistent (i.e., it

[25] Although I have not argued for this claim, one can use the same methods to attribute content to the mental states of the confused. I leave the account of this for future work.

has components that are partial and have different ranges of applicability), then the logic required is even more complicated. For 2 components, one needs 7 semantic values; for 3 components, one needs 14 semantic values; for 4 components, one needs 25 semantic values.

Fourth, inconsistent concepts are fusions of other concepts. I provide no account of how to choose the component concepts for a given inconsistent concept. I have already said that this choice might not be amenable to theory given that it will depend on considerations of overall simplicity and economy. Furthermore, it seems to me that there might be cases where two different sets of component concepts can be equally good candidates for a single inconsistent concept. If so, then which concepts constitute the component concepts of a given inconsistent concept is to some degree indeterminate. Although I do not argue for this claim, I suggest that inconsistent concepts are fusions of *consistent* concepts. That suggestion is an expression of optimism on my part. If it is true, then for any inconsistent concept we encounter, there will be a group of consistent concepts in terms of which it can be explained.

Fifth, the consistent concepts that are components of a given inconsistent concept will be natural candidates for replacing the inconsistent concept in question. Given the considerations I present in Appendix A, for a group of consistent concepts to qualify as replacements for an inconsistent concept, it must be possible to introduce a generic concept for that group. For example, **jadeite** and **nephrite** are the replacement concepts for the confused concept **jade**. To avoid the problems I present in Appendix A, it must be possible to introduce a generic concept, say, ***jade$_g$***, such that x is jade$_g$ if and only if x is jadeite or x is nephrite. The generic concept is handy for people who do not know that jade is confused but are part of a linguistic community in which others do have this knowledge. One can say that such a person employs the generic concept by virtue of the division of linguistic labor.

Sixth, my theory of inconsistent concepts constitutes a much-needed alternative to dialetheism. Dialetheism is the doctrine that some sentences are both true and false, or alternatively, that some contradictions are true. It has several historical precursors, but the contemporary version was proposed by Priest in the late 1970s.[26] One common worry about accounts of inconsistent concepts is that they require dialetheism. Not so. I reject dialetheism for the simple reason that no sentences are both true and false. The following is a quote from David Lewis that summarizes my position exactly:

> The reason we should reject [dialetheism] is simple. No truth does have, and no truth could have, a true negation. Nothing is, and nothing could be, literally both true and false. This we know for certain, and apriori, and without any exception for especially perplexing subject matters. … That may seem dogmatic. And it is: I am affirming the very thesis that Routley and Priest have called into question and – contrary to the rules of debate – I decline to defend it. Further, I concede that it is indefensible against their challenge. They have called so much into question that I have no foothold on undisputed ground. So much the worse for the demand that philosophers always must be ready to defend their theses under the rules of debate, (Lewis 1982: 101).

The reason that dialetheism does not follow from my account of inconsistent concepts is that conceptually inconsistent sentences do not have truth-values. Hence, the sentences that dialethists claim are both true and false have no truth-values on my account.

For example, 'R is a rable and it is not the case that R is a rable' is a sentence that a dialethist might want to call both true and false. It is certainly a contradiction (in the sense that it has the form ⌈p & ~p⌉, but *not* in the sense that it is false in all possible worlds, or in the sense that it is false by virtue of its logical form, or in the sense that any sentence follows from it). However, it has no truth-value according to my theory. Hence, some contradictions are not false. That is a long way from saying that some contradictions are true.

[26] See Priest (1979, 1989, 1998).

6.7 Objections and Replies

Objection 1: The theory of confusion explains confusion in terms of what it is to take or treat a person or expression or concept as confused. However, in the examples cited (mass, jade, etc.) these concepts were confused when no one knew that they were. An adequate theory of confusion has to be able to explain cases of *hidden confusion* like these.

Reply 1: It is a mistake to think that because a theory of X explains X in terms of what it is to treat something as an X, the theory cannot explain cases of hidden X-ness. In particular, the theory of confusion I endorse can explain cases of hidden confusion. For example, prior to 1863, no one knew that jade was a confused concept. However, because of a certain fact (namely, that people prior to 1863 were applying jade to both nephrite and jadeite without distinguishing between the two) it was appropriate to treat the people in question as confused. Unfortunately, no one at the time had this knowledge because they were ignorant of certain empirical facts. Thus, the theory of confusion I endorse handles cases of hidden confusion perfectly well.[27]

Objection 2: Camp's theory of confusion is just as inferentially uncharitable as the ambiguity theory and the theories that attribute truth-values to confused sentences. To continue Camp's example, Fred does not know that he is confused. If we assume that Fred usually reasons according to classical logic, then he will reason according to classical logic when using 'Charlie' as well. According to Camp's theory of confusion, a relevance logic is appropriate for

[27] See Field (1994b, 1998, 2000, 2001g, 2001h), Leeds (1997, 2000), Schiffer (1998), and Akiba (2002a, 2002b) for discussions of issues surrounding hidden conceptual defectiveness. Again, Leport (2004) casts doubt on the historical accuracy of the jade example.

evaluating Fred's arguments that contain 'Charlie'. Some arguments that are valid according to classical logic are invalid according to relevance logic (e.g., those that depend on disjunctive syllogism). Because Fred does not know that he is confused, he will present and endorse some arguments that are valid according to classical logic but invalid according to relevance logic. Thus, Camp's theory of confusion will treat Fred as if he is a poor reasoner. Therefore, it fails to meet its own standard of inferential charity.

Reply 2: Because Fred does not know that he is confused, he will certainly present and endorse arguments that are classically valid but are invalid according to Camp's theory of confusion. Thus, Camp's theory implies that some of his arguments are invalid even though he has every reason to think they are valid. However, that is not the same as treating Fred as if he is irrational. Camp's theory implies that Fred endorses some inference rules that he should not endorse. However, the ambiguity account and the accounts that attribute truth-values to confused sentences imply that Fred does not know how to follow the inference rules he endorses—they treat him as if he makes trivial logical mistakes. There is an important difference between correctly following an inference rule one accepts that happens to be invalid in the context in which one employs it, and incorrectly following an inference rule one accepts.[28] Perfectly reasonable people can disagree about which inference rules are valid in certain circumstances (for examples, look at debates between intuitionists and classical mathematicians, or any discussion between advocates and opponents of non-classical logics). According to Camp's theory, Fred is perfectly reasonable but is correctly following inference rules that are inappropriate for his arguments that contain 'Charlie'. According to the ambiguity account and the accounts that attribute truth-values to Fred's confused sentences, Fred is irrational because he

[28] Camp's theory does imply that the inference rules we ought to endorse are (in part) dictated by the physical environment in which one finds oneself even if one is ignorant of the relevant environmental features.

is not correctly following the inference rules he endorses. Therefore, although Camp's theory is implies that some of the inference rules Fred accepts are invalid, it does not treat Fred as if he is irrational.

Objection 3: There are no inconsistent concepts. All the attempts in this dissertation to define inconsistent concepts fail to define any concept at all. The reason that there are no inconsistent concepts is that interpretation requires one to use the logic one endorses when interpreting another. Thus, it is inappropriate to ever attribute an inconsistent concept to someone, since the interpreter would have to attribute something that defies the logic she endorses.[29]

Reply 3: First, the claim that we interpret others as if they endorse our logical standards is simply false. If it were true then there would be no distinction between criticizing someone for failing to follow her own inference rule and criticizing someone for endorsing the wrong inference rules. It is obvious that there is such a distinction and it plays an important role in philosophical discussions. Second, charity can cut both ways. One might simply introduce an inconsistent concept, begin using it, and describe it as inconsistent. It seems to me that it would be quite difficult to go on interpreting someone who does this as if they had misunderstood their own stipulative definition and their claims about it. Indeed, one might give an account of all the relevant factors in charitable interpretation and present two situations, one in which the weighted sum of all the factors is higher than that of the second, while in the first one attributes an inconsistent concept, but in the second one does not. The point here is that attributing an inconsistent concept is sometimes the most charitable thing to do. No matter what constraints one imposes on charitable interpretation (except of course, a conceptual consistency constraint), there will be situations in which it is more charitable to attribute an inconsistent concept.

[29] One can find a similar objection in Stebins (1992).

Objection 4: Inconsistent concepts are unemployable because every object in the domain in question is in both the application set and the disapplication set of an inconsistent concept. For example, assume that β is a member of the overdetermination set for 'rable' and that α is any other object. Because β is a member of the overdetermination set for 'rable', β is a rable. Hence, either β is a rable or α is a rable. It also follows from the assumption that β is not a rable. Therefore, α is a rable. Analogous reasoning leads to the claim that α is not a rable. Therefore, if the overdetermination set for 'rable' is not empty, then it is the universal set.[30]

Reply 4: According to my theory of inconsistent concepts, this argument is invalid. Assume that we treat 'x is a rable' as confused and that its components are 'x is a table' and 'x is not red'. Since there are only two components, use the 2-component logic with the epistemic values to evaluate arguments involving 'rable'. We substitute 'x is a table' and 'x is not red' alternatively for 'x is a rable' and use the validity condition to evaluate the argument. The result is that it is invalid. This should not come as a surprise because it employs disjunctive syllogism, which is not a valid inference rule in most relevance logics. Note that I am not invoking relevance logic in some ad hoc way just to deal with this objection. It is dictated by my theory of inconsistent concepts. I discuss a different version of this objection to my inconsistency theory of truth in Chapter Seven.

[30] One can find similar objections in Gupta and Belnap (1993). See Chihara (1984) for a discussion as well. This argument is known as *Lewis's argument*. It is unclear who first used it; see Dunn and Restall (2001) and Mares (2004) for discussion.

6.8 CONCLUSION

This chapter completes my theory of inconsistent concepts. In Chapter Four, I presented the outlines of the theory, on which a concept is inconsistent if and only if its constitutive rules for employment are incompatible. I discussed several types of inconsistent concepts and presented several examples. Of particular importance was my commitment to explaining inconsistent concepts in terms of confused concepts. Accordingly, in Chapter Five, I presented my preferred theory of confusion and the logic for confused expressions associated with it. In Chapter Six, I extended Camp's theory of confusion to include confused concepts with more than two components and those that have partial concepts as components. I extended the logic to include conditionals and quantifiers. I also presented Brandom's theory of meaning and used it to construct a pragmatic theory for confusion and a semantic theory for confused expressions. Using this extended theory of confusion as a basis for my theory of inconsistent concepts, I am now ready to apply my theory of inconsistent concepts to truth. That is the topic of the next chapter.

7.0 AN INCONSISTENCY THEORY OF TRUTH

7.1 INTRODUCTION

In this, the final chapter of the dissertation, I apply the theory of inconsistent concepts from Chapters Four, Five, and Six to truth. The goal is to arrive at a theory of truth and a semantic theory for truth that do not generate revenge paradoxes or self-refutation problems; such a theory of truth does not need to be restricted in any way. Thus, the goal is a theory of truth that can serve as the basis for a descriptively correct semantic theory for truth that satisfies the strong internalizability requirement defended in Chapter Two.

In the next two sections of this chapter, I discuss deflationism and partial truth predicates in an effort to sort out several issues before presenting the inconsistency theory of truth. The major issue for an inconsistency theory of the sort I advocate is the choice of components. I advocate an inconsistency theory on which truth has six component concepts. I present them and my reasons for choosing them in the first part of section four. The rest of section four is devoted to a logic, a pragmatic theory, and a semantic theory for truth. That completes the inconsistency theory of truth I advocate for our everyday notion of truth. Of course, I also endorse the replacement policy for inconsistent concepts. Thus, I claim that we should stop using our everyday concept of truth and replace it with one or more consistent concepts. Hence, the inconsistency theory of truth I present is a theory I hope to render obsolete with a change in our linguistic practice. In section five, I present what I take to be an adequate team of replacement concepts for our everyday concept of truth. These replacements turn out to be the six component concepts of truth. It is important to keep in mind that I propose two distinct theories of truth. The first is a descriptive theory on which our concept of truth is an inconsistent concept. The second is a revisionary theory that stipulates how we should change our linguistic practice. I close the chapter by considering some objections.

7.2 Deflationist Truth

Deflationists agree that truth should not be explained in terms of some substantive notion like correspondence, coherence, or utility. Of course, deflationists offer more than this negative claim, but there are at least half a dozen different theories that are currently popular with philosophers who consider themselves deflationists, and certainly several times that many deflationist theories are no longer endorsed by those concerned with truth. I do not intend to provide a detailed discussion of the varieties of deflationism here. I should say that my approach to the nature of truth is broadly deflationist, but I qualify that in several different ways as I present the theory. For now, it means that I do not subscribe to any of the alternative analyses of truth (correspondence, coherence, epistemic, pragmatic), and truth, for me, is aptly called a logical predicate.

According to one type of deflationism, for any sentence, ⟨p⟩, ⟨p⟩ and ⟨⟨p⟩ is true⟩ have the same content. When I say that they have the same content, I mean exactly that—they have the same content. They have all the same semantic features that depend on content. If one holds that biconditionals whose components have the same content are metaphysically necessary, then ⟨⟨p⟩ is true if and only if p⟩ is metaphysically necessary. If propositional attitudes are individuated no more finely than content is individuated, then one believes that p if and only if one believes that ⟨p⟩ is true. If assertions are individuated no more finely than content is individuated, then one asserts that p if and only if one asserts that ⟨p⟩ is true. And so on. Just to avoid confusion, I will use ‘true$_{\text{D}}$’ for a truth predicate with these features.

Although there has been some debate about this in the literature, it seems to me that some have failed to recognize that ‘true$_{\text{D}}$’ is not a “real” predicate. By this I do not mean that ‘true$_{\text{D}}$’ does not behave like a predicate; it does. It has the surface grammar of a predicate. However, sentences of the form ⟨⟨p⟩ is true$_{\text{D}}$⟩ do not attribute any property to ⟨p⟩. The reason? Consider an example. If we disregard the truth rules (i.e., ⟨p⟩ ⊣ ⊢ ⟨⟨p⟩ is true⟩) for a moment, then from ‘my dog is asleep’ nothing follows about ‘my dog is asleep’. Likewise, nothing follows about ‘my dog is asleep’ from ‘‘my dog is asleep’ is true$_{\text{D}}$’. Although the latter’s surface grammar makes it seem like ‘there exists something identical to ‘my dog is asleep’’ follows from it, it does not. The reason is that it does not follow from ‘my dog is asleep’. There is no property that is being attributed to ‘my dog is asleep’ when one asserts ‘‘my dog is asleep’ is true$_{\text{D}}$’. Note that I am not claiming that all philosophers who answer to the term ‘deflationist’ think that ‘true$_{\text{D}}$’ captures our truth predicate in English.

There is an analogous deflationist use of ‘false’, which I label ‘false$_{\text{D}}$’; ⟨it is not the case that p⟩ and ⟨⟨p⟩ is false$_{\text{D}}$⟩ have the same content. The same point holds for ‘false$_{\text{D}}$’; i.e., false$_{\text{D}}$ attributions are not “real” attributions—they are not used to attribute a property. ‘true$_{\text{D}}$’ and ‘false$_{\text{D}}$’ are mere logical devices that one can use to do whatever one might want to do with ⟨p⟩, but without using ⟨p⟩.

There are several different ways of explaining the content of sentences that contain ‘true$_D$’ and ‘false$_D$’, but I prefer an anaphoric account whereby sentences in which these terms occur inherit their content from antecedents. Grover, Camp, and Belnap introduced the prosentential theory of truth in an attempt to explain our everyday concept of truth in terms of an expression that functions anaphorically.[1] I prefer Brandom’s version of the prosentential theory, according to which ‘is true$_D$’ is treated as a prosentence-forming operator.[2] If there is no sentence from which a sentence containing ‘true$_D$’ inherits its content, then the sentence is without content, just like ‘he is a good guy’ would be if ‘he’ was used anaphorically but had no antecedent. It seems to me that one could borrow the distinction between meaning and content from accounts of indexicals and say that such a sentence has a meaning, but no content.[3]

I want to point out that, on this explanation, ‘true$_D$’ and ‘false$_D$’ do not give rise to aletheic paradoxes. Consider the versions of the sentences common to aletheic paradoxes that contain ‘true$_D$’ and ‘false$_D$’.

(1) (1) is false$_D$.

Sentence (1) is a prosentence whose content is a function of the content of its antecedent. Which antecedent? Well, its antecedent is supposed to be (1). Obviously, ‘(1)’ is ambiguous (i.e., there are lots of sentences named ‘(1)’). I assume that in the context at hand, the token of ‘(1)’ in (1) refers to sentence (1) in this chapter and not some other sentence (1). Now that we have determined its antecedent, we can determine its content, which should be a function of the content of its antecedent. There seem to me to be at least two alternatives here. First, we could say that the string of symbols that begins with ‘(1)’ on the fourth line of this paragraph is false,

[1] Grover, Camp, and Belnap (1976).

[2] Brandom (1994).

[3] Note that although I have endorsed Brandom’s version of the prosentential theory of truth here, I have not endorsed his preferred way of dealing with the aletheic paradoxes (Kripke’s semantics).

because it implies that two sentences have the same content even though one is the negation of the other. If so, then the sentence used to introduce the name of the sentence I have been calling sentence (1) failed. Hence, either ‘sentence (1)’ refers to some other sentence or it fails to refer. If it refers to some other sentence, then the content of sentence (1) is a function of the content of that other sentence. If it fails to refer, then it has no content. Recall that we can say that it still has a meaning (just like a sentence that contains an indexical, but is not uttered in a context that is sufficient to determine a content). The other option is to say that sentence (1) fails to have content because it does not inherit content from a sentence that actually has content. Again, it has meaning, but no content. It is akin to ‘it is false’ when ‘it’ is used anaphorically and fails to have an antecedent. I prefer the second alternative, but each implies that the liar paradox is not an issue for a deflationist truth predicate.

A person confronted with sentence (1) might present the following sentences as an argument:

(a) Assume that (1) is true_{D}.

(b) ‘(1) is false_{D}’ is true_{D}.

(c) (1) is false_{D}.

(d) Hence, if (1) is true_{D}, then (1) is $\text{false}_{D.}$

(e) Assume that (1) is $\text{false}_{D.}$

(f) ‘(1) is false_{D}’ is true_{D}.

(g) (1) is true_{D}.

(h) Hence, if (1) is false_{D}, then (1) is true_{D}.

(i) Therefore, (1) is true_{D} if and only if (1) is false_{D}.

However, none of these sentences is truth-valued; they all suffer from the same problem as (1). Hence, they do not constitute an argument because they have no content.

Three further issues deserve comment. First, one cannot use ‘true$_D$’ or ‘false$_D$’ to describe the semantic features of paradoxical sentences. Indeed, one cannot use ‘true$_D$’ or ‘false$_D$’ to describe the semantic features of anything unless the anaphoric antecedent of the sentence in which they occur does so. Second, it will be impossible to determine whether some sentences containing ‘true$_D$’ and ‘false$_D$’ have content. I do not see this as a problem, because in these cases, we would not be able to determine what their content was even if they had it. However, we can always determine their meaning. If, for example, Jack asserts ‘everything Aristotle said the day he died is true$_D$’, then it is impossible for either Jack or anyone who hears him to determine the content of the sentence he uttered. However, if one can determine the antecedent(s) of a sentence that contains ‘true$_D$’ or ‘false$_D$’ and determine their content, then one can determine the content of the sentence in question. Third, I do not think that this variety of deflationism can explain our expression ‘true’ in English for at least two reasons: (i) we sometimes use ‘true’ to attribute a property to truth bearers instead of as a prosentence-forming operator, and (ii) this version of deflationism cannot explain ungrounded but non-paradoxical sentences (i.e., ‘every sentence is either true or false’) because it implies that they are contentless.

7.3 Partial Truth

In this section, I discuss several partial concepts of truth. Partial concepts came up in Chapter Four when I presented several examples of inconsistent concepts. There I gave Soames' example of a partial concept:

(2a) 'smidget' applies to x if x is greater than four feet tall.

(2b) 'smidget' disapplies to x if x is less than two feet tall.[4]

The union of the application set and the disapplication set for a partial concept do not exhaust the domain in question.

We have good reason to believe that truth is a partial concept because we have good reason to believe that: (i) something is true if and only if it is a member of the application set of truth, (ii) something is false if and only if it is a member of the disapplication set for truth, and (iii) there are things that are neither true nor false (e.g., acorns). Of course, that is not a conclusive argument for the claim that truth is a partial concept, and there are philosophers who claim that it is not partial.[5] I do not take issue with them here.

In the first subsection, I discuss the distinction between strong truth and weak truth, which arises for accounts of truth that treat truth as a partial concept. I argue that we need to introduce a third notion of truth, dual truth, in addition to strong and weak truth. In the second subsection, I propose to explain strong truth and dual truth in terms of weak truth. The distinction between these notions of truth and the explanation I offer plays an important role in the account of the components of our inconsistent concept of truth.

[4] Soames (1999).

[5] See Williamson (1997); see Glanzberg (2003) for discussion.

7.3.1 WEAK TRUTH, STRONG TRUTH, AND DUAL TRUTH

In classical logic, all the predicates are completely defined. Once one abandons this assumption, things become more complicated. In languages with partially defined predicates, there are several competing intuitions about the behavior of truth. In particular, if the truth predicate is itself partially defined, then there are two incompatible principles for how to handle truth attributions. For example, assume that p is a sentence and that it is a truth-value gap. What is the truth-status of 'p is true'? On one intuition, 'p is true' should have the same truth status as p; hence, 'p is true' should be a gap as well. On the other intuition, 'p is true' says of p that it is true, but p is not true—it is a gap; hence, 'p is true' should be false. A concept of truth that conforms to the first intuition is a weak concept of truth, while one that obeys the second is a strong concept of truth (this distinction surfaced in Chapter Two during the discussion of Kripke's semantic theory for truth). Yablo seems to have been the first to draw this distinction explicitly:

> To call a true sentence true is to say something true, and to call an untrue sentence true is to say something false; this is what is meant by the assertion that truth is strong. On a competing conception of truth, which may be dubbed the weak conception, the statement that ϕ is true simply inherits ϕ's truth-status, whatever it may be. Thus, if ϕ is neither true nor false, then to call it true is, on the weak conception, to say something neither true nor false; and if ϕ is for some reason both true and false, then to call it true is to say something itself both true and false, (Yablo 1985: 301).

Since Yablo presented his distinction, it has become a staple of discussions of the aletheic paradoxes and the nature of truth.

Strong truth and weak truth agree on the sentences that have truth values; it is only on the truth-value gaps that they differ:

(Weak Truth) If p is true, then 'p is true' is true.
If p is false, then 'p is true' is false.

If p is a gap, then ‘p is true’ is a gap.

(Strong Truth) If p is true, then ‘p is true’ is true.
If p is false, then ‘p is true’ is false.
If p is a gap, then ‘p is true’ is false.

One can define weak falsity and strong falsity in analogous ways. Before moving on, I want to discuss this distinction a bit more. It is surprisingly subtle.

The first thing to notice is that both weak truth and strong truth are partial concepts. That is, both concepts admit of truth-value gaps. I have already assumed that our everyday concept of truth is partial on the set of objects because most of us feel strongly that non-linguistic objects are not true and not false. Moreover, our everyday concept of truth is partial on the set of sentences as well because most of us feel strongly that imperatives and interrogatives are not true and not false. Some philosophers claim that our everyday concept of truth is completely defined on the declarative sentences; that is, they claim that every declarative sentence is either true or false. This claim is also known as the principle of bivalence. There are plenty of debates about this principle, and I am not going to enter into them here. I want to point out that one can draw different distinctions between weak truth and strong truth depending on how one draws the line between their ranges of applicability and their ranges of inapplicability. If both are completely defined on the set of declarative sentences, then they differ only on the non-declarative sentences. That is not very interesting only because humans rarely call non-declarative sentences true or false. To get an interesting distinction, one must assume that one or both are partial on the set of declarative sentences. Some philosophers have argued that truth is not only partial on the set of declarative sentences, but on the set of truth-sentences as well. That is, they claim that some sentences that contain truth predicates are not true and not false. Indeed, Kripke’s semantic theory implies that the liar sentence is one of these.

The next thing to notice is that in the definitions (Weak Truth) and (Strong Truth), the words ‘true’, ‘false’, and ‘gap’ occur. One might wonder which type of predicates these are. Is ‘true’ in these definitions a weak truth predicate or a strong truth predicate? Surprisingly, there are no discussions of which I am aware that address this question. Once we begin to offer answers to it, we find that there are a good many ways to draw the distinction between weak truth and strong truth.

It seems obvious to me that one should be able to treat the ‘true’ in the definition of weak truth as a weak truth predicate, and one should be able to treat the ‘true’ in the definition of strong truth as a strong truth predicate. Given that these two concepts are distinct, one must also keep track of the different types of gaps (i.e., weak gaps and strong gaps). A sentence is a weak gap iff it is in the range of inapplicability for weak truth. A sentence is a strong gap iff it is in the range of inapplicability for strong truth. Things become even more complex when considering mixed truth attributions (e.g., ‘‘p is strong true’ is weak false’).

Because the potential for confusion and misunderstanding is already high, for the rest of this discussion I refrain from using unmodified ‘true’, ‘false’, or ‘gap’. Instead, I use ‘WT’ for weak truth, ‘ST’ for strong truth, ‘WF’ for weak falsity, ‘SF’ for strong falsity, ‘WG’ for weak gaphood, and ‘SG’ for strong gaphood. In addition, when defining the relations between these notions, I use set theoretic terminology instead of using these notions in both the definiens and the definiendum. I use Arial bold type for the corresponding sets (i.e., ‘**WT**’, ‘**ST**’, ‘**WF**’, ‘**SF**’, ‘**WG**’, and ‘**SG**’).[6] The following are the principles governing weak truth, weak falsity, strong truth, and strong falsity predicates:

If p ∈ **WT**, then ‘p is WT’ ∈ **WT** and ‘p is WT’ ∈ **ST**.
If p ∈ **WF**, then ‘p is WT’ ∈ **WF** and ‘p is WT’ ∈ **SF**.

[6] ‘p is WT’ and ‘p ∈ **WT**’ are different claims. The first can be a gap, but the second cannot.

If p ∈ **WG**, then ‘p is WT’ ∈ **WG** and ‘p is WT’ ∈ **SF**.
If p ∈ **ST**, then ‘p is WT’ ∈ **WT** and ‘p is WT’ ∈ **ST**.
If p ∈ **SF**, then ‘p is WT’ ∈ **WF** and ‘p is WT’ ∈ **SF**.
If p ∈ **SG**, then ‘p is WT’ ∈ **WG** and ‘p is WT’ ∈ **SF**.

If p ∈ **WT**, then ‘p is WF’ ∈ **WF** and ‘p is WF’ ∈ **SF**.
If p ∈ **WF**, then ‘p is WF’ ∈ **WT** and ‘p is WF’ ∈ **ST**.
If p ∈ **WG**, then ‘p is WF’ ∈ **WG** and ‘p is WF’ ∈ **SF**.
If p ∈ **ST**, then ‘p is WF’ ∈ **WF** and ‘p is WF’ ∈ **SF**.
If p ∈ **SF**, then ‘p is WF’ ∈ **WT** and ‘p is WF’ ∈ **ST**.
If p ∈ **SG**, then ‘p is WF’ ∈ **WG** and ‘p is WF’ ∈ **SF**.

If p ∈ **WT**, then ‘p is ST’ ∈ **WT** and ‘p is ST’ ∈ **ST**.
If p ∈ **WF**, then ‘p is ST’ ∈ **WF** and ‘p is ST’ ∈ **SF**.
If p ∈ **WG**, then ‘p is ST’ ∈ **WG** and ‘p is ST’ ∈ **SF**.
If p ∈ **ST**, then ‘p is ST’ ∈ **WT** and ‘p is ST’ ∈ **ST**.
If p ∈ **SF**, then ‘p is ST’ ∈ **WF** and ‘p is ST’ ∈ **SF**.
If p ∈ **SG**, then ‘p is ST’ ∈ **WG** and ‘p is ST’ ∈ **SF**.

If p ∈ **WT**, then ‘p is SF’ ∈ **WF** and ‘p is SF’ ∈ **SF**.
If p ∈ **WF**, then ‘p is SF’ ∈ **WT** and ‘p is SF’ ∈ **ST**.
If p ∈ **WG**, then ‘p is SF’ ∈ **WG** and ‘p is SF’ ∈ **SF**.
If p ∈ **ST**, then ‘p is SF’ ∈ **WF** and ‘p is SF’ ∈ **SF**.
If p ∈ **SF**, then ‘p is SF’ ∈ **WT** and ‘p is SF’ ∈ **ST**.
If p ∈ **SG**, then ‘p is SF’ ∈ **WG** and ‘p is SF’ ∈ **SF**.

Notice that they differ when p is a member of **WG** or **SG**. Notice also that no strong truth or strong falsity attributions are members of **SG**, but some are members of **WG**. Thus, the ranges of applicability for weak truth and strong truth are different. Some sentences that are weak truth gaps are not strong truth gaps.

Given that their ranges of applicability are different, one issue is whether they are both partially defined on the declarative sentences. If strong truth is completely defined on the declarative sentences, then the only strong truth gaps are non-declarative sentences, which is not very interesting. I assume that they are both partially defined on the declarative sentences. If one desires an account on which strong truth is completely defined on the declarative sentences, then that will be easy to accommodate.

Note that the application set of weak truth (**WT**) and the application set of strong truth (**ST**) are identical. That should seem odd. Usually when one distinguishes between a strong version and a weak version of a concept, it means that the application set of the strong concept is a proper subset of the application set of the weak concept. That is not the case with weak truth and strong truth. One could, of course, define two notions of truth so that the application set of one is a proper subset of the application set of the other (and one can even name them 'weak truth' and 'strong truth'), but my point is that one need not do so in order to satisfy the definitions given above. Moreover, provided that both concepts of truth are partial, one would have to draw a weak truth/strong truth distinction for each one.

One might object that weak truth and strong truth do have different application sets for the following reason. Let p be a gap. If 'true' expresses weak truth, then 'p is not true' is a gap as well. However, if 'true' expresses strong truth, then 'p is not true' is true. Thus, 'p is not true' is in the range of inapplicability for weak truth and in the application set for strong truth. Hence, they have different application sets.

My reply is that one must distinguish between weak truth and strong truth throughout the argument. If p is a weak gap, then 'p is not weak true' is a weak gap as well, while 'p is not strong true' is strong true. Thus, if we are treating 'true' in 'p is not true' as ambiguous, then it is correct that this sentence will be in the range of inapplicability for weak truth and in the application set for strong truth, but the meaning of 'true' in these two sentences differs. It is synonymous with 'weak true' in one and with 'strong true' in the other. Thus, the objection is like saying that 'First National is a bank' is both true and false because on one reading of 'bank' it is true and on another it is false. Once one distinguishes between the two notions of truth, the argument is invalid.

In situations where the distinction between weak truth and strong truth is appropriate, it is helpful to have another concept of truth as well. This notion of truth has not received any attention in the literature as far as I can tell; for lack of a better term, I call it *dual truth*. One can define it in the following way:

(Dual Truth) If p is true, then ‘p is true’ is true.
If p is false, then ‘p is true’ is false.
If p is a gap, then ‘p is true’ is true.

Because we treat truth as designated and we treat falsity and gaphood as undesignated, we have little use for this notion of truth. That probably accounts for the lack of attention. However, we do have a use for the accompanying notion of falsity. There are situations when one would like to employ a notion of falsity, but neither weak falsity nor strong falsity will do. In these cases, one needs dual falsity. If p is a sentence to which one wants to attribute some notion of falsity, but one does not want both p and $\ulcorner$~p$\urcorner$ to be this kind of false, and one wants to utter a sentence that is going to be either this kind of false or this kind of true, then one wants to attribute dual falsity to p. The following are the principles governing dual truth (‘DT’) attributions and dual falsity (‘DF’) attributions:

If p ∈ **WT**, then ‘p is DT’ ∈ **WT**, and ‘p is DT’ ∈ **ST**, and ‘p is DT’ ∈ **DT**.
If p ∈ **WF**, then ‘p is DT’ ∈ **WF**, and ‘p is DT’ ∈ **SF**, and ‘p is DT’ ∈ **DF**.
If p ∈ **WG**, then ‘p is DT’ ∈ **WG**, and ‘p is DT’ ∈ **SF**, and ‘p is DT’ ∈ **DT**.
If p ∈ **ST**, then ‘p is DT’ ∈ **WT**, and ‘p is DT’ ∈ **ST**, and ‘p is DT’ ∈ **DT**.
If p ∈ **SF**, then ‘p is DT’ ∈ **WF**, and ‘p is DT’ ∈ **SF**, and ‘p is DT’ ∈ **DF**.
If p ∈ **SG**, then ‘p is DT’ ∈ **WG**, and ‘p is DT’ ∈ **SF**, and ‘p is DT’ ∈ **DT**.
If p ∈ **DT**, then ‘p is DT’ ∈ **WT**, and ‘p is DT’ ∈ **ST**, and ‘p is DT’ ∈ **DT**.
If p ∈ **DF**, then ‘p is DT’ ∈ **WF**, and ‘p is DT’ ∈ **SF**, and ‘p is DT’ ∈ **DF**.
If p ∈ **DG**, then ‘p is DT’ ∈ **WG**, and ‘p is DT’ ∈ **SF**, and ‘p is DT’ ∈ **DT**.

If p ∈ **WT**, then ‘p is DF’ ∈ **WF**, and ‘p is DF’ ∈ **SF**, and ‘p is DF’ ∈ **DF**.
If p ∈ **WF**, then ‘p is DF’ ∈ **WT**, and ‘p is DF’ ∈ **ST**, and ‘p is DF’ ∈ **DT**.
If p ∈ **WG**, then ‘p is DF’ ∈ **WG**, and ‘p is DF’ ∈ **SF**, and ‘p is DF’ ∈ **DT**.
If p ∈ **ST**, then ‘p is DF’ ∈ **WF**, and ‘p is DF’ ∈ **SF**, and ‘p is DF’ ∈ **DF**.
If p ∈ **SF**, then ‘p is DF’ ∈ **WT**, and ‘p is DF’ ∈ **ST**, and ‘p is DF’ ∈ **DT**.

If p $\in$ **SG**, then 'p is DF' $\in$ **WG**, and 'p is DF' $\in$ **SF**, and 'p is DF' $\in$ **DT**.
If p $\in$ **DT**, then 'p is DF' $\in$ **WF**, and 'p is DF' $\in$ **SF**, and 'p is DF' $\in$ **DF**.
If p $\in$ **DF**, then 'p is DF' $\in$ **WT**, and 'p is DF' $\in$ **ST**, and 'p is DF' $\in$ **DT**.
If p $\in$ **DG**, then 'p is DF' $\in$ **WG**, and 'p is DF' $\in$ **SF**, and 'p is DF' $\in$ **DT**.

The relations between dual truth and the other two types of truth are fairly straightforward. **WT** and **ST** are identical. **WF** and **DF** are identical. **DG** and **SG** are identical. All the members of **WG** – **DG** (which is identical to **WG** – **SG**) are members of **DT** and members of **SF**. Any truth attribution to any type of truth-value gap is in **WG** $\cap$ **SF** $\cap$ **DT**.

One attributes weak truth if one wants one's sentence to have the same weak truth-status as the target. One attributes strong truth if one wants one's sentence to be in **ST** only if the target is in **ST**; otherwise one's sentence in **SF**. One attributes strong falsity if one wants one's sentence to be in **ST** only if the target is in **WF**; otherwise, one's sentence is in **SF**. A sentence attributing strong truth or strong falsity to a sentence in **WG** is in **SF**. One attributes dual truth if one wants one's sentence to be in **DT** if the target is in either **WG** or **WT**. One attributes dual falsity if one wants one's sentence to be in **DT** only if the target is in **WF**; otherwise, one's sentence is in **DF**. A sentence attributing dual truth or dual falsity to a member of **WG** is in **DT**.

One might find the following to be more helpful definitions of weak truth, strong truth, and dual truth for name-predicate sentences.

(Weak Truth) For any predicate 'H' and any object α:
If $\alpha \in H_E$, then 'α is H' $\in$ **WT**.
If $\alpha \in H_A$, then 'α is H' $\in$ **WF**.
If $\alpha \in H_I$, then 'α is H' $\in$ **WG**.

(Strong Truth) For any *truth* predicate 'H' and any *declarative sentence* α:
If $\alpha \in H_E$, then 'α is H' $\in$ **ST**.
If $\alpha \in H_A$, then 'α is H' $\in$ **SF**.
If $\alpha \in H_I$, then 'α is H' $\in$ **SF**.

For any *other* predicate 'H' and any *declarative sentence* α:
If $\alpha \in H_E$, then 'α is H' $\in$ **ST**.

If $\alpha \in H_A$, then ‘α is H’ $\in$ **SF**.
If $\alpha \in H_I$, then ‘α is H’ $\in$ **SG**.

(Dual Truth) For any *truth* predicate ‘H’ and any *declarative sentence* α:
If $\alpha \in H_E$, then ‘α is H’ $\in$ **DT**.
If $\alpha \in H_A$, then ‘α is H’ $\in$ **DF**.
If $\alpha \in H_I$, then ‘α is H’ $\in$ **DT**.

For any *other* predicate ‘H’ and any *declarative sentence* α:
If $\alpha \in H_E$, then ‘α is H’ $\in$ **DT**.
If $\alpha \in H_A$, then ‘α is H’ $\in$ **DF**.
If $\alpha \in H_I$, then ‘α is H’ $\in$ **DG**.

In the above definitions, ‘H_E’ ‘H_A’ ‘H_I’ designate the application set of ‘H’, the disapplication set of ‘H’, and the range of inapplicability for ‘H’, respectively. For example, assume that ‘red’ is the predicate in question and ‘1’ is the name in question. Assume as well that the number 1 is in the range of inapplicability for ‘red’. Thus, ‘1 is red’ is a member of **WG**, **SG**, and **DG**; that is, it is a weak gap, a strong gap, and a dual gap. All three truth predicates agree on whether sentences that do *not* contain truth predicates are gaps. They disagree on how to classify sentences that *do* contain truth predicates. To continue the example, let ‘p’ be the name of ‘1 is red’. The sentence ‘p is weak true’ is a member of **WG**, but it is not a member of **SG** or **DG**. Instead, it is a member of **SF** and **DT**. ‘p is weak false’ is also a member of **WG**, **SF**, and **DT**. ‘p is strong true’ and ‘p is strong false’ are members of **WG**, **SF**, and **DT**. On the above definitions, all sentences can be assigned three different types of truth-statuses: a weak truth-status, a strong truth-status, and a dual truth-status. For any sentence p that does not contain a truth predicate (or falsity predicate), the weak truth-status, the strong truth-status, and a dual truth-status of p are the same. Thus, one can speak of the truth-status (simpliciter) of sentences that do not contain truth predicates. The truth-statuses of sentences that do contain truth predicates will differ.[7] (Notice that if one is of the opinion that strong truth and dual truth are completely defined on the declarative sentences, then one can alter the above definitions so that what would have been the members of **SG** and **DG** are members of **SF** and **ST**, respectively.)

7.3.2 Partial Truth Predicates

In this subsection, I explain strong truth and dual truth in terms of weak truth. At this point, I would like to argue that strong truth and dual truth can be defined in terms of weak truth with the help of exclusion negation. One way to argue for such a claim would be to define a 3-valued language L with the usual sentential operators and exclusion negation and extend it to L′ by adding a weak truth predicate (i.e., ‘WT-in-L′’) to L′ and prove that one can define a strong truth predicate (i.e., ‘ST-in-L′ ’) and a dual truth predicate (i.e., ‘DT-in-L′’) in L′. Unfortunately, there are several problems with this strategy.

The first problem is that the best a formal setting could do is show that one can define ‘strong truth-in-L’ and ‘dual truth-in-L’ in terms of ‘weak truth-in-L’, where ‘L’ is the name of a formal language. That is not good enough for my purposes, because I am interested in the unrestricted weak, strong, and dual truth predicates. Still, the formal account would provide some evidence that the same relation holds between the unrestricted versions.

The second problem is more pressing. Tarski showed that no language in which the classical inference rules are valid and which has the capacity to represent its own syntactic

[7] One can say that the weak truth predicate is *tolerant* in the sense that it does not force ‘α is H’ to be either true or false when α is outside the range of applicability for ‘H’. The strong truth predicate and the dual truth predicate are tolerant with respect to the sentences that do not contain truth predicates, but they are intolerant with respect to the sentences that contain truth predicates (think of a lawyer forcing a witness to answer a “yes/no” question whose presuppositions fail). One can define different notions of weak truth, strong truth, and dual truth that have different ranges of tolerance and intolerance, and there are some interesting results in this area. Unfortunately, they will have to wait for some other occasion.

features can contain its own truth predicate without being inconsistent.[8] That result does not affect my proposal because L′ would not be a classical language. There are many proposals for how to construct non-classical languages that contain their own truth predicates. The most influential is Kripke's theory of truth. Kripke begins with a classical first-order language and adds a partial truth predicate. He constructs a sequence of languages such that in each one, more and more sentences of the language are placed in either the extension or the anti-extension of the truth predicate that belongs to that language. Kripke shows that this procedure eventually reaches a language in which no more sentences are added to either. A language with this property is called a *fixed point*. For Kripke's procedure to result in a fixed point, the sequence of languages must be *monotonic*; i.e., if L_2 is a language that comes later in the sequence than L_1, and p is in the extension (anti-extension) of 'true' in L_1, then p is in the extension (anti-extension) of 'true' in L_2.[9] Roughly, once a sentence gets a truth-value, it keeps that truth-value throughout the construction. One can show that if the sequence of languages is monotonic, then they do not contain any non-monotonic sentential operators. A sentential operator is *monotonic* if and only if changing one of the components from a gap to a truth-value does not change the compound from one truth-value to the other or from a truth-value to a truth-value gap. Roughly, one can "fill in" the gaps with truth-values without affecting the truth-value of the compound so long as it has a truth-value. In a sentential compound whose sentential operator is monotonic, the truth-value of the compound (if it has one) is determined by the truth-values of the components that have truth-values.

Kripke's procedure is unavailable because L′ would contain exclusion negation, which is a non-monotonic sentential operator. Gupta and Martin proved that a language with a weak

[8] Tarski (1933).

[9] Kripke (1975). The only differences between the languages is the extension and anti-extension of 'true'; thus, it makes sense to think of p as belonging to all the languages in the sequence.

Kleene scheme can contain a non-monotonic sentential operator and still reach a fixed point, but their procedure will not work for a language with exclusion negation.[10] I know of no procedure for adding a truth predicate to a language with truth-value gaps that contains exclusion negation.

There is a good reason for this problem. I have argued in Chapter Three that weak truth is inconsistent. If that is right, then any language that has the capacity to represent its own syntax, that validates the truth rules, and that has exclusion negation (which is used to formulate revenge paradoxes) is inconsistent. One could use the account I offer of how to interpret such languages, but that would beg the question because it presupposes the claim that strong truth and dual truth can be defined in terms of weak truth. Instead of providing a formal language and a formal definition, I offer an informal account of weak truth, strong truth, and dual truth.

I begin by noting that the three notions of truth agree on the sentences that do not contain truth predicates. I assume that one has an adequate definition of weak truth in the sense that one knows its application set (**WT**), its disapplication set (**WF**), and its range of inapplicability (**WG**). One then defines the application set of strong truth (**ST**) so that it is identical to **WT**. The disapplication set of strong truth (**SF**) is the union of **WF** and the set of truth attributions that belong to **WG**. One defines dual truth by stipulating that its application set (**DT**) is the union of **WT** and the set of truth attributions that belong to **WG**. The disapplication set of dual truth (**DF**) is identical to **WF**.

One can define the sentential operators in terms of weak truth and then explain the way strong truth and dual truth interact with sentential operators by appeal to the above relations between them. For example, the following are the principles that describe the relations between weak truth and three types of negation (*choice negation* (~), *exclusion negation* (¬), and *inclusion negation* (−)):

[10] Gupta and Martin (1984).

If $p \in \mathbf{WT}$, then $\ulcorner {\sim}\, p \urcorner \in \mathbf{WT}$.
If $p \in \mathbf{WF}$, then $\ulcorner {\sim}\, p \urcorner \in \mathbf{WF}$.
If $p \in \mathbf{WG}$, then $\ulcorner {\sim}\, p \urcorner \in \mathbf{WG}$.
If $p \in \mathbf{WT}$, then $\ulcorner \neg\, p \urcorner \in \mathbf{WT}$.
If $p \in \mathbf{WF}$, then $\ulcorner \neg\, p \urcorner \in \mathbf{WF}$.
If $p \in \mathbf{WG}$, then $\ulcorner \neg\, p \urcorner \in \mathbf{WT}$.
If $p \in \mathbf{WT}$, then $\ulcorner -\, p \urcorner \in \mathbf{WT}$.
If $p \in \mathbf{WF}$, then $\ulcorner -\, p \urcorner \in \mathbf{WF}$.
If $p \in \mathbf{WG}$, then $\ulcorner -\, p \urcorner \in \mathbf{WF}$.

One issue is how the negation operators interact with strong truth and dual truth. Given that strong truth and dual truth are identical to weak truth for sentences that do not contain truth predicates, for sentences that do not contain truth predicates, strong truth and dual truth behave just like weak truth. For truth attributions, one defines strong truth and dual truth by appeal to weak truth. For example, let r be a member of **WG**, **SG**, and **DG**. Thus, r contains no truth predicate. ‘r is weak true’ and ‘r is weak false’ are both members of **SF**. To determine whether ‘~ (r is weak true)’ is a member of **ST** or **SF**, one first determines whether ‘~ (r is weak true)’ is a member of **WT**, **WF**, or **WG**. By the above definition, ‘~ (r is weak true)’ is a member of **WG**. Given the relation between truth sentences of **WG** and **SF**, ‘~ (r is weak true)’ is a member of **SF**. The other sentential operators are defined on strong truth and dual truth in the same way. The result is that the sentential operators are weak truth functional, but not strong truth functional or dual truth functional.

This explanation of strong truth in terms of weak truth is a striking departure from the assumption that the application set of strong truth is a proper subset of the application set of weak truth. On the contrary, the only difference between weak truth and strong truth is their disapplication sets. The disapplication set of strong truth contains all the truth-attributions that are in either the disapplication set of weak truth or the range of inapplicability of weak truth.

The diagrams below illustrate the difference between weak truth, strong truth, and dual truth. Circles represent the set of truth-attributions of a given language:

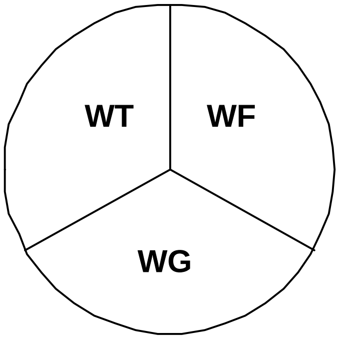


Figure 7.1 (Weak Truth)

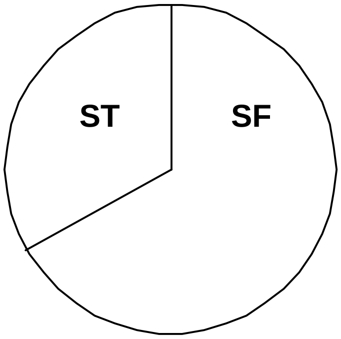


Figure 7.2 (Strong Truth)

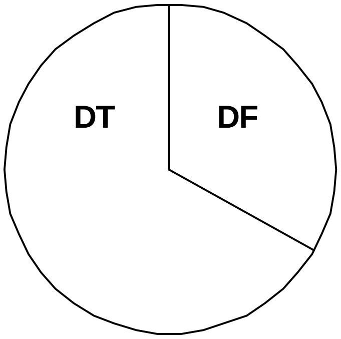


Figure 7.3 (Dual Truth)

One final issue deserves comment: the status(es) of truth tellers. A truth teller is a sentence like:

(τ) (τ) is true.

Truth tellers are less perplexing than liars, because they do not engender paradoxes. However, they are still troubling because one can consistently assign either truth or falsity to them. Just as one can show that a contradiction follows from the claim that a liar is true or false, one can show that a tautology follows from the claim that a truth teller is either true or false. Furthermore, if one assigns (τ) to **WG**, one can construct a new “revenge” truth teller that can consistently be assigned either truth, falsity, or gaphood. It seems to me that just as the liars and revenge liars show that weak truth is inconsistent, the truth tellers and revenge truth tellers show that weak truth is partial on the truth-sentences. I assume that any truth teller should be treated as a weak gap. Notice that all the truth tellers will be members of **SF** and **DT** as well.[11]

[11] This result vindicates Yablo’s intuitions regarding truth tellers; see Yablo (1985, 2003).

The point of this section is to sort out the issues surrounding weak truth, strong truth, and dual truth in an effort to get clear on the different notions of truth one requires when one treats truth as a partial concept. It seems obvious to me that there is no good reason to think that weak truth, strong truth, or dual truth is our everyday notion of truth. Rather, I claim that we should think of our everyday concept of truth as a confused concept whose components are weak truth, strong truth, and dual truth. However, that claim alone does not constitute a solution to the problems associated with the liar paradox. The problem is that we can still generate liar-type paradoxes with weak truth (I discussed them at length in Chapters Two and Three), and, since strong truth and dual truth are defined in terms of weak truth, they give rise to paradoxes as well. Thus, nothing in this section should be taken as a solution to the liar paradox. Rather, I take it to have justified the claim that the choice of components for truth should respect the distinction between weak truth, strong truth, and dual truth.

7.4 Inconsistent Truth

In this section, I offer a theory of truth on which truth is an inconsistent concept, and I propose a logic, a pragmatic theory, and a semantic theory for truth that are based on this theory of truth. There should not be any surprises in the general outlines of my account—it is just the theory of inconsistent concepts developed in Chapters Four, Five, and Six applied to truth. The logic, the pragmatic theory, and the semantic theory for truth are versions of those presented in Chapters Five and Six. Given my claims that: (i) truth is an inconsistent concept, (ii) inconsistent concepts should be explained in terms of confused concepts, and (iii) confused concepts should be explained by appeal to their components, the most pressing issue is the choice of components for truth. That is the topic of the first subsection.

7.4.1 Components of Truth

On one interpretation, Tarski held an inconsistency view of truth. He actually claimed that natural languages are inconsistent, but he thought that they are inconsistent, in part, because of the truth predicates they contain.[12] Other notable inconsistency theorists include Chihara, Priest, Yablo, McGee, and Eklund, all of whom I discuss at other points in this dissertation.[13] Chihara and Yablo provide ways of thinking about concepts with inconsistent definitions (the account I presented in Chapter Four owes much to their views) and McGee simply states that the concept of truth (or our understanding of it) is inconsistent and needs to be replaced. McGee concentrates on his favored replacement instead of on how best to understand an inconsistent concept of truth. Eklund presents a modified version of supervaluation semantics for inconsistent concepts and applies it to truth. Priest advocates dialetheism, which is the doctrine that some contradictions are true; the aletheic paradoxes supply one justification for his view. My account differs considerably from each of these. (See Appendix E for discussion of each of these approaches and a comparison to the theory I advocate.)

On the account I offer, the claim that truth is an inconsistent concept is cashed out as the claim that truth is a confused concept. On the theory of confused concepts I endorse, a confused concept is a fusion of two or more concepts, which are called the *components* of the confused concept. These components play two roles in the theory: (i) they are used in the logic, the

[12] Tarski (1944).

[13] Chihara (1973, 1979, 1984), Priest (1979, 1987), Yablo (1985, 1993a, 1993b), McGee (1991), and Eklund (2002). A number of other philosophers have made remarks that suggest they are sympathetic to the inconsistency view, including Mates (1981), Parsons (1984), Barwise and Etchemendy (1987), Tappenden (1994), and Orlilia (2000).

pragmatic theory, and the semantic theory for the confused concept, and (ii) they serve as replacements for the confused concept. Thus, the logic, the pragmatic theory, and the semantic theory for the confused concept are used to interpret discourse in which the confused concept is expressed; however, the linguistic practice to which the discourse in question belongs should be changed so that the confused concept is no longer employed and replacement concepts are used in its place. Thus, the hope is that the linguistic practice changes so that the logic, the pragmatic theory, and the semantic theory for the confused concept are no longer needed (except to interpret portions of discourse from the old linguistic practice).

As I have suggested, the everyday notion of truth is inconsistent, and it is inconsistent in more than one way. That is, given that our everyday concept of truth is an amalgam of weak truth, strong truth, and partial truth, it is range-inconsistent. Furthermore, given that weak truth is itself inconsistent, our everyday concept of truth is application-inconsistent as well. The components I suggest reflect this complexity. I can give no assurances that I have chosen the optimum components. Indeed, there are certainly other ways of breaking up truth into components, and some of these might work better than the one I propose.

I suggested in the previous section that one could define weak truth, strong truth, and dual truth in several different ways depending on how one draws the line between their ranges of applicability and their ranges of inapplicability. Instead of worrying about all these options, I concentrate on those I defined in the previous section. I assume that weak truth and strong truth have the same application set, that weak truth and dual truth have the same disapplication set, and that strong truth and dual truth have the same range of applicability. Finally, I assume that their range of applicability is some proper subset of the set of declarative sentences (as I have emphasized, this assumption can be dropped if one is committed to a version of bivalence).

Truth is range-inconsistent because some sentences are within the range of applicability for strong truth and dual truth, but belong to the range of inapplicability for weak truth. Truth is application-inconsistent because the intersection of the application set and disapplication set for weak truth is not empty (it contains the paradoxical sentences). Hence, treating truth as a partial concept (and distinguishing weak, strong, and dual truth predicates) does not solve the aletheic paradoxes. That was one of the lessons of Chapter Three. When a concept is both application-inconsistent and range-inconsistent, I find it helpful to deal with the range-inconsistency first. I have distinguished between the three mentioned truth predicates to do so. Each of these truth predicates is application-inconsistent because weak truth is application-inconsistent and strong truth and dual truth are defined in terms of weak truth. Addressing the application-inconsistency of weak truth will solve the problem with strong truth and dual truth as well.

Before presenting my suggestion for the components of truth, I want to consider and reject several alternative suggestions. One suggestion involves the Tarskian truth predicates. That is, one can treat truth as if it is an amalgam of all the Tarskian concepts of truth. There are several difficulties with this proposal. First, it is exceedingly difficult to characterize *all* the Tarskian truth predicates because the task requires a transfinite hierarchy, which poses its own problems (see Appendix A for a discussion of some of these). Another problem is that the Tarskian truth predicates do not serve as adequate replacements for the concept of truth (again, Appendix A has the details). Finally, this suggestion does nothing to remedy the range-inconsistency present in our concept of truth because all the Tarskian truth predicates are completely defined.

A second suggestion appeals to the claim, common in deflationist theories of truth, that the set of T-sentences of a language implicitly define its truth predicate or serve as its theory of truth. Of course, the problem with such a view is that the set of T-sentences for any natural language is inconsistent. It might seem natural to treat the T-sentences as meaning-constitutive sentences for truth and explain its inconsistency in terms of the inconsistency of the set of T-sentences. I argue in Appendix E that such theories are inadequate to explain inconsistent concepts because they do not respect the distinction between concept possession and concept employment.

However, one might still want to take a cue from the T-sentences when determining the components for weak truth. For example, it might seem plausible to say that weak truth is an amalgam of two concepts, each of which is governed by a subset of the T-sentences such that the two sets are disjoint and mutually exhaustive of the set of T-sentences. Unfortunately, no such strategy can work. No matter how one divides the set of T-sentences into two subsets, at least one will still be inconsistent. Consider the T-sentence for the standard liar (i.e., '(λ) is false'): '(λ) is false' is true iff (λ) is false. If (λ) is in the language in question, then this T-sentence will be a member of the set of T-sentences for that language, and this T-sentence is a contradiction in classical logic. No matter how one splits up the set of T-sentences, whichever subset contains this T-sentence will be inconsistent. Thus, to use the T-sentences to define the components of weak truth, one must eliminate some of them. Of course, the problem is which ones to eliminate? McGee has shown that there are indefinitely many distinct maximally consistent sets of T-sentences, that none of them are axiomatizable, and that the only sentences that belong to the intersection of all of them are T-sentences for truth tellers.[14] I do not see the set of T-sentences as a good place to look for components of weak truth.[15]

[14] McGee (1991).

[15] Another serious problem with this approach is that whether a sentence is paradoxical can depend on factors beyond its semantic features; one can construct a sentence whose paradoxicality depends on most anything. If one's approach to the liar paradox involves eliminating the T-sentences for paradoxical sentences of the language, then the

I propose to look to the truth rules, in particular, the ascending and descending weak truth rules. The ascending and descending weak truth rules are: ⟨p⟩ follows from ⟨⟨p⟩ is weak true⟩ and ⟨⟨p⟩ is weak true⟩ follows from ⟨p⟩. Recall that these were the basis for the behavior of the weak truth predicate. That is, ⟨⟨p⟩ is weak true⟩ has the same truth status as ⟨p⟩. I believe that these rules are also the reason for the inconsistency of weak truth. I propose to split weak truth into two concepts, one based on one of the weak truth rules, the other based on the other rule. I call them *ascending weak truth* and *descending weak truth*. Ascending weak truth, 'AWT', is governed by an ascending weak truth rule: ⟨⟨p⟩ is AWT⟩ follows from ⟨p⟩. Descending weak truth, 'DWT', is governed by a descending weak truth rule: ⟨p⟩ follows from ⟨⟨p⟩ is DWT⟩.

It seems to me that one should stipulate restricted versions of the other rule for both types of weak truth. Ascending weak truth obeys a restricted version of the descending weak truth rule, and descending weak truth obeys a restricted version of the ascending weak truth rule. The restricted versions should hold for non-pathological sentences and fail for pathological ones. One defines the pathological sentences by appeal to the paradoxical sentences for weak truth. If p is a paradoxical sentence for weak truth in the sense that one can derive that p is both weak true and not_E weak true using the weak truth rules[16], then the result of substituting 'AWT' for all the occurrences of 'weak true' and 'AWF' for all the occurrences of 'weak false' is a pathological sentence, and the result of substituting 'DWT' for all the occurrences of 'weak true' and 'DWF' for all the occurrences of 'weak false' is a pathological sentence. For example, the following are pathological sentences:

(α) (α) is AWF.

concepts of truth that result will depend on unrelated empirical facts. See Appendix B for a full discussion of this problem.

[16] On my view, there is no non-circular way of defining paradoxicality. Certainly, there is no syntactic or semantic definition of paradoxicality (I argue for this claim in Appendix B). The best we can hope for is that a sentence is paradoxical iff one can derive that it is both true and not true from principles governing weak truth.

(δ) (δ) is DWF.

They result from substituting ‘AWF’ and ‘DWF’ for ‘weak false’ in the weak liar. Notice that pathological sentences are *not* paradoxical—they do not give rise to paradoxes.

The two types of weak are truth very similar. Of course, they are different concepts with different application sets and disapplication sets. Most important, both are consistent concepts. The following are principles governing ascending weak truth (AWT) and descending weak truth (DWT):

If p ∈ **DWT**, then ‘p is AWT’ ∈ **DWT**.
If p ∈ **DWT**, then ‘p is DWT’ ∈ **DWT**.
If p ∈ **DWT**, then ‘p is AWF’ ∈ **AWF**.
If p ∈ **DWT**, then ‘p is DWF’ ∈ **AWF**.

If p ∈ **AWF**, then ‘p is AWT’ ∈ **AWF**.
If p ∈ **AWF**, then ‘p is DWT’ ∈ **AWF**.
If p ∈ **AWF**, then ‘p is AWF’ ∈ **DWT**.
If p ∈ **AWF**, then ‘p is DWF’ ∈ **DWT**.

If p ∈ **DWF**, then ‘p is AWT’ ∈ **AWF**.
If p ∈ **DWF**, then ‘p is DWF’ ∈ **DWT**.

If p ∈ **AWT**, then ‘p is AWT’ ∈ **DWT**.
If p ∈ **AWT**, then ‘p is DWF’ ∈ **AWF**.

The intersection of **AWT** and **DWF** is not empty. In fact, (α) and (δ) are members of this set. In other words, (α) and (δ) are both ascending weak true and descending weak false. **DWT** and **AWF** contain non-pathological sentences (which include all the truth-apt sentences without truth predicates). **AWF** is a proper subset of **DWF**, and **DWT** is a proper subset of **AWT**. Aside from being partial, ‘AWT’ and ‘DWT’ behave just like any other predicates. In the previous section, I used a weak Kleene scheme for sentential operators (other than exclusion negation and inclusion negation), but I prefer to accommodate a range of views on how best to handle partial concepts.

Once ascending weak truth and descending weak truth are available, one can define strong and dual versions of each according to the account given in the previous section. The result is a team of six truth concepts that are all consistent and can be used in the logic, the pragmatic theory, and the semantic theory for the inconsistent concept of truth.

To recapitulate, here is my proposal for how to divide up the concept of truth. Truth is range-inconsistent. To address this inconsistency, I distinguish between weak truth, strong truth, and dual truth. Weak truth, strong truth, and dual truth are application-inconsistent. To address this inconsistency, I distinguish between ascending weak truth and descending weak truth. I define ascending strong truth and ascending dual truth in terms of ascending weak truth, and I define descending strong truth and descending dual truth in terms of descending weak truth. That makes for six component concepts for truth.

7.4.2 ALETHEIC LOGIC

I find it helpful to present an example of a community of language users that employ an inconsistent concept of truth. The *Aletheians* (ăl′-ĭ-thē′-ənz) are my model community. It is also helpful to be able to describe the activities of an interpreter who attempts to understand and describe the linguistic practice of the Aletheians. I call the interpreter in my story, *Mojo*. In this subsection, I describe the way the Aletheians use 'true'. Just to avoid confusion, I use the term 'true$_A$' for their truth predicate.

It should be obvious that I want Mojo's interpretation of the Aletheians' concept of truth to apply to our own concept of truth. Hence, their linguistic practice should be as much like ours as possible. Accordingly, they speak English, or at least a language that is very much like the American English of 2004 (my idiolect, noon GMT on Jan. 1). I do not want to be burdened with language identity problems so I stipulate that all the Aletheians speak exactly the same language. There is no difference between the idiolects and the communal language. The Aletheians' language (I call it *Aletheian*) is a bit different from English because their linguistic practice does not display a division of linguistic labor or a speaker's reference/term reference distinction. They do use indexicals, demonstratives, pronouns, and other context dependent expressions. They also use the usual modal, normative, moral, epistemic, logical, temporal, mental, and semantic vocabulary. They experience reference failure, presupposition failure, category mistakes, confusion, identity ignorance[17], and the other common types of defective discourse.

As far as their use of 'true_A' is concerned, I want it to express an inconsistent concept for them. However, I am not going to specify the way they use it in advance. To do so would be to lay out a theory before I have my example set up. They use 'true_A' just like we use 'true'; they use 'false_A' just like we use 'false', except that there is no issue about the legitimacy of their aletheic paradoxes. The application conditions for 'true_A' determine that it both applies and disapplies to some sentences—the paradoxical ones. They are unaware that their concept of truth_A is inconsistent, but they are aware of the paradoxes. They believe that paradoxicality is the result of their god's anger. That is, they believe that their god has caused them to make mistakes in their use of 'true_A', which result in calling some sentences both true_A and not true_A. They believe that their god does this to punish them for their transgressions. Although they cannot figure out where the supposed mistakes took place, they think that looking into the matter too much will further anger the god and bring more conceptual wrath. Hence, they rarely reflect on their uses of 'true_A'.

[17] Identity ignorance is the phenomenon that drives Frege's puzzle and Kripke's puzzle. One can think of it as inverse confusion and use an inverse confusion logic, pragmatics, and semantics to sort out the issues Frege and Kripke raise; but this is a story for another occasion.

Although they are perplexed about the phenomenon of paradoxicality, they effectively employ the containment strategy as far as possible. That is, once it is discovered that a sentence is paradoxical, the Aletheians think of it as tainted and so they stop using it and stop talking about it. They are not proactive about the containment strategy. They do not check each sentence they intend to use before using it for paradoxicality. Instead, they only worry about paradoxicality once someone shows that a given sentence is paradoxical. Not surprisingly, this approach rarely leads to problems in everyday communication. The Aletheians are not much for semantic theorizing so the paradoxes do not bother them the way they bother us.

Mojo has the thankless task of constructing a logic, a pragmatic theory, and a semantic theory for Aletheian. It should not come as a surprise that Mojo agrees with me on the relevant issues. In particular, he accepts the theory of inconsistent concepts presented in Chapters Four, Five, and Six, and he intends to use it to arrive at the theories he needs.

Let me emphasize the fact that Mojo is not a radical interpreter. For one, he understands Aletheian. Nor is he restricted to the materials of the radical interpreter. Mojo has access to any non-linguistic fact he needs. He can also determine the mental states of the Aletheians with incredible accuracy—not that it will help him much. He needs to come up with a way of interpreting the Aletheians' statements involving 'true_A'. First, he will have to sort out several issues surrounding the truth predicate itself. His main goal is to determine the component concepts he will use in constructing his logic, pragmatics, and semantics. If truth is inconsistent, and inconsistent concepts are confused, then truth is confused. If truth is confused, then there are component concepts that have been fused together into truth. To use the theory of inconsistent concepts, Mojo must decide on a set of component concepts. This is no small task. I have already indicated that I have no algorithm for how to decide on component concepts. Instead of worrying about how Mojo goes about it, I assume that he has already made the choice and decided on using the six component concepts I suggest.

In this subsection, I concentrate on the logic that is appropriate for the Aletheians' truth-sentences. Mojo first needs to assemble a crack team of experts on the various component concepts of truth. He will have to rely on them when determining the aletheic logic to use for evaluating the Aletheians' arguments involving truth. He then determines which Aletheian sentences contain 'true_A' as a predicate. He then compiles a list of all of them and a list of all the all the possible Aletheian arguments.

Which logic should Mojo use? Given that there are six components of truth, he will need a 6-component logic; given that the components of truth are partially defined and have different ranges of applicability, he will need a partial 6-component logic. The query values are: Y, N, ?, and G. A partial 6-component logic has more response values than I care to list, but they are analogous to the response values for the other partial n-component logics. For a partial 6-component logic, the response values are grouped into semantic values in the usual way; it turns out that a partial 6-component logic has 63 semantic values. Mojo assigns each of the sentences on his list one of these semantic values by asking his experts for query values for each, tabulating the response value for each, and deriving the semantic value for each. For example, if the response values for a sentence p are Y, N, Y, ?, N, and G, then the semantic value for p is YNG. For each of the logically compound sentences on his list, he uses the rules for the sentential operators given in Chapter Six. Given that he is using a partial 6-component logic, negation, disjunction, and conjunction are straightforward; however, the conditional is not. He cannot rely on a link between his logic and a standard relevance logic (like R). Instead, he can introduce a non-embeddable conditional based on the validity criterion.

Using the rules for validity in partial n-component logics from Chapter Six, Mojo determines whether each of the arguments on his list is valid or invalid. Given that he has to use a partial n-component logic, there is no general principle associating the valid arguments from his logic with valid arguments in a standard relevance logic (like R). Despite these difficulties and the complexity involved, he has a workable logic with which he can evaluate the relevant arguments of the Aletheians.

7.4.3 ALETHEIC PRAGMATICS

After using the aletheic logic to evaluate the Aletheians' arguments, Mojo can use the confusion pragmatics to keep score on them and determine which of their assertions are warranted. There are no surprises here. The only difference between what Mojo does for the Aletheians and what Ginger did to Fred in Chapter Five is that Mojo can use the 'G' value to rule out warrant right away. If one of the Aletheians utters a sentence and that sentence is assigned a 'G' under the 63-valued aletheic logic, then that utterance does not count as warranted. Otherwise, Mojo can use the logic to assign entitlements to the Aletheians.

7.4.4 ALETHEIC SEMANTICS

Again, no surprises. Mojo uses the confusion semantics to determine freestanding assertional contents and freestanding ingredient contents for the Aletheians' sentences that contain 'true$_A$'. His procedure is no different than Ginger's in Chapter Five.

7.5 FROM INCONSISTENT TRUTH TO CONSISTENT TRUTH

Mojo has a procedure by which to interpret the Aletheians' use of 'true_A'. I do not want to suggest that he is obligated to try to replace their inconsistent concept of truth with the six component concepts. However, I want to consider what would happen if the Aletheians discovered that their concept of truth is inconsistent and decided to implement the six components as a team of replacements.

I do not want to get into the procedure by which they determine that their truth predicate is inconsistent or their transformation to consistency. One problem that surfaces after their transformation is complete is that the theory of truth I have constructed is no longer correct—the Aletheians no longer employ an inconsistent concept of truth. No matter. If they accept the suggested replacements, then the semantics appropriate for them will work for the Aletheians as well. I did not give a semantics for the six component predicates, but I do not anticipate any problems with them. Of course, they are partial concepts so the semantic theory for them will have to treat them accordingly.

One point about the sentences that express the inconsistent concept of truth is that they are all gaps. That is, they are all in the ranges of inapplicability for ascending weak truth and descending weak truth. That fact reflects the conclusion from Chapter Five that sentences that express defective concepts do not have truth values.

How do the replacements fare in actual conversation? Among the component predicates, the ascending weak truth predicate and the descending weak truth predicate are primary (since the other four can be defined in terms of them). Can the Aletheians tell when to use one over the other in conversation? Certainly, the choice of one of the weak truth predicates versus one of the strong ones or one of the dual ones is easy enough to make. But can they effectively choose between the two weak ones? Only on the pathological sentences will the two weak truth predicates differ. A pathological sentence is descending weak false and ascending weak true. Consider the standard liar reasoning. The argument from ⟨p⟩ to ⟨⟨p⟩ is ascending weak true⟩ is valid but the argument from ⟨⟨p⟩ is ascending weak true⟩ to ⟨p⟩ is invalid for pathological sentences. On the other hand, the argument from ⟨p⟩ to ⟨⟨p⟩ is descending weak true⟩ is invalid for pathological sentences, but the argument from ⟨⟨p⟩ is descending weak true⟩ to ⟨p⟩ is valid.

Still, it might seem that if an Aletheian is presented with a sentence and decides to evaluate it with respect to one of the types of weak truth, she needs to know whether it is pathological before she can choose between ascending weak truth and descending weak truth. Thus, it seems that the Aletheians will have trouble using these truth predicates—they seem to impose unreasonable epistemic demands on their users.

The Aletheian does not need to know whether the target of his attribution is pathological to decide which weak truth predicate to use. The extension of descending weak truth is a proper subset of the extension of ascending weak truth. The extension of descending weak truth contains no pathological sentences, while the extension of ascending weak truth contains all the pathological sentences. Thus, descending weak truth is stronger than ascending weak truth. A sentence can be both ascending weak true and descending weak false, but no sentence is both descending weak true and either descending or ascending weak false. Thus, a descending weak truth attribution is, in some sense, more careful than an ascending weak truth attribution. If the Aletheian in question (let us call him *Alex*) wants to make sure that the sentence he utters is not both ascending weak true and descending weak false, then he should use descending weak truth. If he does not care, then he should use ascending weak truth. Using ascending weak truth gives one a better chance of uttering a sentence that is ascending weak true. However, it also opens one up to uttering a sentence that is both ascending weak true and descending weak false. If the idea of uttering a sentence that is both ascending weak true and descending weak false is unsavory to Alex, then he should use descending weak truth. If he uses descending weak truth, then his sentence is guaranteed to be either descending weak true or ascending weak false (or a weak gap). Therefore, the Aletheians can decide which weak truth predicate to use on a given occasion despite the fact that they are often not in a position to determine whether a given sentence is pathological.

7.6 Objections and Replies

Objection 1: On the view presented in this chapter, all sentences with truth predicates are gaps. That is radically counterintuitive.

Reply 1: First, a clarification—all the sentences that express the inconsistent concept of truth are gaps (they are ascending and descending weak gaps, ascending and descending strong gaps, and ascending and descending dual gaps). However, many of the sentences that express one of the replacement concepts of truth are not gaps. Moreover, the vast majority of the sentences that express the inconsistent concept of truth can be reformulated with one of the replacement concepts such that they have ascending and descending weak truth-values. I agree that the account I offer of the sentences that express the inconsistent concept of truth might seem counterintuitive at first, but it makes sense once one admits that truth is a defective concept. For example, it does not seem counterintuitive to say that all the sentences that express the Newtonian concept of mass are gaps. Given the very minor changes required to retire the inconsistent concept of truth and begin using the team of consistent concepts, one can begin uttering truth-valued sentences containing truth predicates in no time. Compared to the revolution that had to occur before we could start using an appropriate concept of mass, this one is painless.

The objector might respond by saying: given that on the theory of truth (our everyday concept) I accept, sentences that express this concept are truth-valueless (for any concept of truth), this theory does a poor job of explaining our intuitions about our everyday concept of truth (e.g., that many sentences that express this concept have truth-values); thus, the theory I advocate does not explain our concept of truth as well as many of its competitors. In the face of the objection, I plead guilty that the theory I have developed does not respect all of our intuitions about truth. However, because our intuitions about truth are inconsistent, only an inconsistent theory could respect them all. Everyone who works on truth should admit that much. As for the explanatory adequacy of the theory I propose, it satisfies the strong internalizability requirement and it is one of the only consistent theories to do so. Many of the other theories of truth (all the various gap theories, revision theories, contextualist theories, etc.) fail to meet the internalizability requirement because they face either revenge paradoxes or self-refutation problems. I argued at length in Chapters One and Two that, from an explanatory standpoint, the internalizability requirements are not negotiable. Thus, when understood properly, none of these theories even purports to explain the concept of truth—never mind how good a job they do. Thus, when it comes to the game of explaining our concept of truth (not some restricted version of it), the theory I offer is at least a player—that is more than can be said for these other theories. Of course, there are several theories that purport to meet the internalizability requirement; I claim that only one other meets it (Eklund's theory) and that the theory I offer is superior to it (see Appendix D and Appendix E for discussion of this issue).

Here is a related worry: even if some philosophers accept this account (and that is a big 'if'), no one else is going accept the change to our linguistic practice I recommend. I agree, but that does not pose a problem for me because the division of linguistic labor will effectively force the change on everyone provided that most people will defer to experts concerning difficult matters of usage. For example, even if I do not know that there are two concepts of mass, I will still defer to experts on the topic of mass. So long as I am willing to do this, it makes sense to interpret my word 'mass' as a generic term for mass, which is synonymous with 'relativistic mass or proper mass'. The same goes for the replacement concepts of truth.

Consider a further worry: no philosophers are going to accept the changes I propose because, given the centrality of truth, they force more changes in other areas of philosophy (e.g., assertibility, meaning, reference, predication, knowledge, justification, validity, proof, soundness, completeness, etc.) My response is that accepting the replacement concepts of truth will have some effects on mathematical and philosophical logic; however, they are minor and the costs of change are more than outweighed by the benefits that come from a consistent conceptual scheme. Given that, on a deflationist account of truth, one can explain most of these other concepts without a substantive appeal to truth, they should survive unscathed. I suppose that some changes will have to take place to accommodate the presence of defective concepts (e.g., truth-conditional theories of meaning will have to be amended or abandoned), but given that these changes need to be made anyway, there is no added burden for the theory of truth.

Objection 2: Inconsistent concepts are unusable because every object in the domain in question is in the overdetermination set for an inconsistent concept.

Reply 2: Recall that I responded to a general form of this objection in section 6.6.3 of Chapter Six. I showed that the argument on which this objection depends is invalid in the logic for inconsistent concepts. However, there is a more insidious version of this objection I must address. Instead of the disjunctive syllogism argument I considered before, Gupta and Belnap use a variant of Curry's paradox to argue that if truth in particular is an inconsistent concept, then it is unusable. Hence, according to Gupta and Belnap, if truth is inconsistent, then the rules for using truth dictate that every sentence is paradoxical.[18] Curry's paradox concerns the following sentence:

(γ) If (γ) is true, then 0 = 1.

The following argument is used to derive the paradox (obviously, any sentence could be used in place of '0 = 1'):

(a) (γ) is true.

(b) 'if (γ) is true, then 0 = 1' is true. (Substitution)

(c) if (γ) is true, then 0 = 1. (Descending)

(e) 0 = 1. (Modus Ponens)

(f) if (γ) is true, then 0 = 1. (Conditional Proof)

(g) 'if (γ) is true, then 0 = 1' is true. (Ascending)

(h) (γ) is true. (Substitution)

(h) 0 = 1. (Modus Ponens)

Given that the aletheic logic I presented to interpret the Aletheians' truth predicate does not include an embeddable conditional, it is difficult to give a particular evaluation of this argument. However, because the argument relies on using both the ascending and descending truth rules for

[18] Gupta and Belnap (1993: 13-15).

a paradoxical sentence, it will turn out to be invalid no matter what conditional is added to the logic.[19]

Objection 3: I have said several times that truth is an inconsistent concept and that once one recognizes that a concept one possesses is inconsistent, one should refrain from employing it. Yet, I have employed it throughout this dissertation.

Reply 3: I have three explanations for my actions. First, if the only way to inform someone that a certain concept is inconsistent is to employ it, then it seems to me that one ought to go ahead and employ it. For example, to show that 'rable' is inconsistent I might find a red table, point at it, and say 'this is rable' and 'this is not rable'. That is a very special use of the concept, and it seems to me that it should be legitimate.[20] Some of my uses of truth in this dissertation are of this sort. Second, an assertion of 'truth is an inconsistent concept' is not an employment of the concept of truth. It is an employment of the concept of the concept of truth, which need not be inconsistent even though truth is. Most of my uses of 'truth' in this dissertation are of this sort. Third, most of the dissertation is independent of my views on truth. I want my claims about internalizability, inconsistent concepts, confused concepts, and partial truth to be acceptable even to those who do not share my view that truth is an inconsistent concept. My use of the word 'true' can be filled out in many different ways. One who agrees with my view that truth is an inconsistent concept can think of those uses of 'true' as expressing ascending weak truth and can think of my uses of 'false' as expressing descending weak falsity.

Here is a deeper worry: I have to employ the inconsistent concept of truth to construct the theory of inconsistent concepts, the semantic theory for truth, or to introduce the replacement concepts; hence, the account is circular. I disagree. None of these theories depend for their

[19] Furthermore, there is good reason to believe that (γ) should be a weak gap, because it is a variant of the truth-teller (τ).

[20] See Tappenden (1994) for discussion.

construction on the inconsistent concept of truth. For example, the semantic theory for truth uses epistemically interpreted semantic values instead of truth values. Even the notion of validity to which the logic appeals is explained in terms of profitability, not truth, and, as deflationists have argued for the past several decades, there is no good reason to think that profitability must be explained in terms of truth.[21]

Objection 4: Why should we think that this view does not give rise to revenge paradoxes or self-refutation problems?

Reply 4: It is difficult to show that the replacement concepts do not give rise to revenge paradoxes. Of course, I could give a model-theoretic semantics for a formal language with predicates that behave somewhat like the replacement concepts and present a relative consistency proof, but that would not show much. First, it would show only that language-specific versions of the replacement concepts are consistent so long as set theory is consistent. However, language-specific versions of the replacement concepts are inadequate as replacements (see Appendix A for more on this issue). Second, even if I were to present such a proof, that would not show that the replacement concepts (even the language-specific versions of them) do not give rise to revenge paradoxes. For example, Field presents just such a proof for his theory of truth, but it *does* give rise to revenge paradoxes. Field avoids the revenge paradoxes by excluding the relevant linguistic resources from the formal language he considers.

I can say that if the account of the revenge paradoxes I gave in Chapter Three is accurate, then we have very good reason to believe that the replacement concepts do not engender revenge paradoxes. On that explanation, revenge paradoxes occur for accounts of truth that respect the truth rules and try to contain the paradoxical sentences by assigning them a defectiveness status.

[21] See Putnam (1967), Horwich (1978, 1990), Leeds (1978, 1995), Soames (1984), Williams (1986), Field (1986, 1994), Resnik (1990), Devitt (1991), Wright (1992), Gupta (1993), and Clark (1997) for discussion.

The revenge paradox occurs if the set of defective sentences is too small. On my account of inconsistent truth, all the sentences that express the concept of truth are defective. Thus, it does not give rise to a revenge paradox. On the theory of the replacement concepts, no concept of truth obeys the truth rules for the relevant sentences. Thus, they do not give rise to revenge paradoxes.

As for self-refutation problems, the theory of the inconsistent concept of truth implies that the inconsistent concept of truth has an empty extension and an empty antiextension because all inconsistent concepts do. However, this theory is not self-refuting because it does not purport to exemplify this concept of truth. Thus, it is not in the same position as a gap theory that implies that its consequences are gaps. On the theory of the inconsistent concept of truth, this concept should not be employed at all. The theory of the replacement concepts does purport to be true—descending weak true. And it is. The theory does not imply that any of its consequences are gappy. Thus, it does not suffer from a self-refutation problem.

Objection 5: In reply 1, I mentioned a generic truth predicate, and I claimed that someone who is unaware that there are several different kinds of truth could be interpreted as using this generic predicate, provided that he is willing to defer to the experts of his linguistic community on matters related to truth. Doesn't the generic concept give rise to a revenge paradox?

Reply 5: No. There are several ways to think of the generic predicate, and none of them give rise to a revenge paradox. It seems to me that the best way to treat someone in this situation is to interpret his use of 'true' as a weak ascending truth predicate and his use of 'false' as a weak descending falsity predicate (if a sentence is weak descending true, then it is weak ascending true, and if a sentence is weak ascending false, then it is weak descending false). That interpretation maximizes the chances that his truth and falsity attributions will be ascending weak true. Of course, he will be using ‘true’ and ‘false’ under the false assumption that there is nothing to which both of these predicates apply, but I do not see that as a problem. Someone who believes that there is nothing to which both ‘egg-layer’ and ‘mammal’ apply can still use them effectively.

Objection 6: Assume that Mojo is interpreting an Aletheian, Alex, and Alex is employing the inconsistent concept of truth. On the aletheic logic used by Mojo, Alex’s argument for the contradiction associated with the liar paradox is valid. Of course, the aletheic logic keeps the contradiction from spreading, but the argument to the contradiction is valid nevertheless. If Mojo is using the aletheic logic together with the pragmatic theory to determine which of Alex’s sentences is assertible, then the same sentence (i.e., the liar) is both assertible and not assertible. Therefore, this approach to the liar paradox just exports the problem to accounts of assertibility.

Reply 6: It is not the case that Alex’s argument for the contradiction is valid on the Aletheic logic. On the contrary, the logic implies that the truth rules are valid for non-paradoxical sentences, but they are not both valid for paradoxical sentences. Thus, Alex’s argument is invalid. Nevertheless, when dealing with defective concepts, one is forced to admit two notions of assertibility and keep score appropriately. For the most familiar notion of assertibility, a sentence is assertible iff the asserter is entitled to assert it. Of course, from the point of view of someone who recognizes that a certain concept is defective, sentences that express this defective concept are not assertible because defective concepts should not be employed. However, there is a derivative notion of assertibility that one can employ even in cases where a defective concept is being employed. On this notion of assertibility, a sentence is assertible iff, given the information at the asserter’s disposal, he has good reason to believe that he is entitled to assert it. On the account of inconsistent concepts I offer, paradoxical sentences might be assertible in this derivative sense. There is no sense in which this result poses a problem for the notion of assertibility, and there is no sense in which the account I offer trades in problems with truth for problems with assertibility.

Objection 7: Gupta and Belnap dispute the coherence of my position because they argue that inconsistency theorists are disguised context dependence theorists:

> The more thoroughly inconsistent a set of conventions is, the less guidance it provides to behavior and the more context dependent are its applications. Hence, if the Inconsistency View is correct, we should expect our use of 'true' in a given context to depend on how we choose, in that context, to interpret and apply the inconsistent conventions governing it. As a result, we should expect our use of 'true', even in reference to context-independent sentences, to be highly dependent on context. This expectation, however, is not borne out in our ordinary uses of 'true', (Gupta and Belnap 1993: 16).

Thus, if truth were an inconsistent concept, it would display context-dependence, but it does not.

Reply 7: I agree that our use of 'true' in reference to context-independent sentences is not dependent on context. However, according to inconsistency theorists, our concept of truth is fine for most sentences, only paradoxical sentences pose a problem and they are few and far between. We do not see contextual variation because there are very few sentences to which our concept of truth directs us to both apply it and disapply it. Furthermore, even for the paradoxical sentences, we do not see contextual variation. We do not affirm them in some contexts and reject them in others. Thus, it is ascending weak false that inconsistent concepts necessarily display context dependence.

Proponents of context dependent approaches say that we do see contextual variation for paradoxical sentences. In one context, we say that sentence (1) is paradoxical, where sentence (1) is 'Sentence (1) is not true', and in another we say that it is not paradoxical, indeed, it is true, because it accurately describes the semantic features of sentence (1); i.e., it says that sentence (1) is not true and it is not, because it is paradoxical. Hence, we do see context dependence for the paradoxical sentences; the extension of ‘true’ differs from context to context. In the first context the extension of ‘true’ includes sentence (1), while in the second context, it does not. That is why sentence (1) is paradoxical in the first context and it is not paradoxical in the second. I have two replies. First, if truth displays this sort of context dependence, then it is not an inconsistent concept. Thus, these considerations do not support Gupta and Belnap’s claim that inconsistent concepts inevitably display context dependence. Second, there is no good reason to think that truth displays this sort of context dependence and there are plenty of good reasons to think that it does not; I enumerate them in Appendix B.

Objection 8: One can construct sentences that are paradoxical for weak truth by virtue of empirical facts that are independent of their syntactic and semantic features (I call this the *riskiness thesis* in Appendix B). Likewise, one can construct a sentence that is paradoxical in one context and non-paradoxical in another even though it has the same syntactic and semantic features in each context. Because ascending weak truth and descending weak truth are defined in terms of paradoxicality, and paradoxicality varies from context to context, their extensions vary from context to context. Thus, the replacement concepts do display context dependence. Therefore, this is a contextualist theory of truth.

Reply 8: I endorse the riskiness thesis and the claim that whether a sentence token is paradoxical can vary from context to context. However, it does not follow that the replacement concepts are context dependent. For a given sentence token, p, whether p is ascending or descending weak true does depend on contextual features, but this sort of context dependence we find in our everyday notion of truth anyway. If I write a sentence token on a card and whip it out from time to time throughout the day, on some occasions it will be considered true and on others it will be considered false. All this reasoning shows is that sentence tokens are not appropriate truth bearers. When presenting theories of truth, it is often helpful to assume that sentences or sentence tokens are truth bearers, but, strictly speaking, they are not. I do not advocate propositions as truth bearers for reasons presented in Appendix B. Rather, it seems to me that the best choice of truth bearers is pairs of sentence tokens and contexts appropriate to determine paradoxicality status. Given that this is the right choice of truth bearers, the replacement concepts are not context dependent—they have the same extensions and anti-extensions (of truth bearers, not sentence tokens) in all contexts.

Objection 9: Truth is indispensable. Hence, it cannot be replaced. The following passage from Leeds expresses this intuition well:

> A theory of truth should not allow us to say, for example, 'A, but 'A' is false;' or 'A, but in one sense 'A' is false, in another true,' etc. Speaking for myself, I would take this as an obvious desideratum for any account of truth: indeed, I think that if we were somehow to become persuaded to use the word 'true' in ways that conflicted with the T-sentences, we would immediately – so important are the disquotational uses of truth in our own language – invent an additional notion of truth – say truth* – that conformed to them; under such circumstances, I think one might as well say that we had never abandoned the T-sentences after all: we had merely decided to rename truth 'truth*' and use the word 'true' to mean something else, (Leeds 1995: 8).

Given that truth is indispensable, there is no chance of replacing it with the replacement concepts of truth.

Reply 9: This view depends on the assumption that truth is consistent. Once it is recognized to be an inconsistent concept, this view is no longer so appealing. Of course, we could introduce an inconsistent concept of truth that obeys all the T-sentences (as Leeds suggests), but it is not going to be as useful as Leeds thinks. Indeed, once it is recognized as an inconsistent concept, it is also recognized as a concept that should not be employed at all. Of course, Leeds' intuition is that without a truth predicate that behaves as the disquotationalists claim, our language would be expressively impoverished. I agree. However, one need not introduce an inconsistent concept of truth to receive the benefits of a deflationist truth predicate. Indeed, I argued in section 7.1 that deflationist truth predicates do not pose any risk of paradox at all. Thus, one could introduce a deflationist truth predicate into the language without much concern—so long as one recognizes that it is distinct from the inconsistent concept of truth that gives rise to the liar paradox.

Here is a deeper worry: why shouldn't we just continue using our inconsistent concept of truth in the way prescribed by the theory of truth presented above (i.e., as governed by a partial 6-component relevance logic) instead of introducing the replacement concepts? My answer: one cannot just use a concept one knows to be inconsistent in accord with the rules for it. The theory of inconsistent truth outlined above is a theory for how to interpret someone who is using this concept, it is not a theory for how one *should* use this concept. The theory for how one should use this concept is very simple: it has no extension and no anti-extension, so one should not apply or disapply it to anything. Once one accepts this, one should see the pointlessness of introducing such a concept or retaining the one we already have.

Objection 10: In Chapter Three, I argued against approaches to the liar paradox that posit some hidden semantic feature of truth (e.g., ambiguity, context dependence, circularity, intensionality) to dismiss the paradox. In particular, these theories are implausible in their account of truth and they do nothing to solve the liar paradox because one could reintroduce the paradox by constructing a new concept that does not have this hidden semantic feature. However, the theory presented above has just this flavor—it posits a hidden semantic feature (i.e., inconsistency) in an effort to dismiss the paradox. Thus, it is just as bad as the other theories that do this.

Reply 10: I agree that on the theory I offer, our concept of truth has a hidden semantic feature—conceptual inconsistency. However, this theory is quite different from the others listed above that I criticized in Chapter Three. First, there is no reason to think that we should have been able to detect the inconsistency in our concept before we acquired the logical sophistication we have today. For inconsistent concepts, it makes sense to think that they can be used for centuries without recognizing that they are inconsistent. That is not the case for context dependent concepts or ambiguous words—we should be able to detect those right away. I do not know if a concept can display hidden circularity, but it seems to me that it can.

Second, the other approaches that posit a hidden semantic feature fail because they try to vindicate truth—they try to show that the liar paradox is not genuine. Not mine. I argue that the liar paradox is genuine—it has its source in our concept of truth. Thus, the theory I offer does constitute a real solution to the paradox. It is a multi-part solution. We need to begin by understanding why the paradox arises—because truth is an inconsistent concept. Once we do that, we need a theory that explains such concepts and allows us to make sense of those who employ them. Finally, we need to know how to change our linguistic practice so that the paradox no longer shows up. The objection that I presented in Chapter Three—that the theories in question do not really solve the paradox because one could just introduce a concept that does not have the hidden semantic feature—does not apply to the theory I present. Of course, one could introduce a concept that does not have this hidden semantic feature (i.e., inconsistency), but that will not reintroduce the paradox, and if one decided to introduce a new inconsistent concept, then the account I give explains what we should do about it.

Objection 11: My view is that ‘true’ has a certain content and that sentences containing truth predicates have contents that are determined, in part, by the content of the truth predicate.

On my view, the content of sentences in which truth predicates occur is different from the content most people think they have, and the content of a truth predicate is different from the content most people think it has. However, on the only respectable view of content, truth predicates and the sentences in which they occur have their contents because of the ways they are used. How can the truth predicate have a content that is different from the content people think it has if it is the actions of those people that are responsible for it having its content? Furthermore, if my view is correct, then no one uses truth predicates according to the real rules that govern them. How can that be? Does it even make sense to say that everyone uses it improperly?

Reply 11: I agree that a condition on any theory of linguistic phenomena is that linguistic items (e.g., sentences, words) have the contents they have because of the ways they are used. The theory I offer respects this condition, but one must alter one's account of use somewhat to accommodate uses of predicates that express inconsistent concepts. For an inconsistent concept, there are three sets of rules. First, there are the inconsistent rules that those who employ the concept without knowing that it is inconsistent try to follow. Second, there are the rules for how it is to be used that are prescribed by the logic, the pragmatic theory, and the semantic theory for the inconsistent concept. These rules are pertinent for those who know that the concept is inconsistent, but are faced with the problem of interpreting those who use it without knowing it is inconsistent. Third, there are the rules for how it is to be used by those who know that it is inconsistent. On these rules, it should not be used at all (except maybe in demonstrations of its inconsistency). The first set of rules describe how it should be used by *everyone* from the point of view of *someone who does not know it is inconsistent*; the second set describe how it should be used by *those who do not know it is inconsistent* from the point of view of *those who know that it is inconsistent*; and the third set describe how it should be used by *those who know it is* *inconsistent* from the point of view of *those who know it is inconsistent*. Almost everyone uses it according to the first set of rules, and a few enlightened ones use it according to the third set of rules; however, no one uses it according to the second set of rules. It is the second set of rules that are used to determine its content for *those who do not know it is inconsistent* from the point of view of *those who do know it is inconsistent*.

Without this distinction between the three sets of rules, there would be no distinction between concept possession and concept employment. The pragmatic theory allows one to keep track of these distinctions. That should not be surprising. When I am talking with someone who I know endorses different inference rules, I can keep track of whether they are following their own rules and whether they are following my rules. In the case of someone who employs the inconsistent concept of truth, they are committed to following the rules for the use of the concept, but they think that it has different rules than it actually does. Moreover, it has the rules it actually has (i.e., it should not be used) because of the rules everyone thinks it has (i.e., it should be used according to principles that turn out to be inconsistent). The commitments undertaken by those who employ the concept outrun the commitments they acknowledge. This case is no different from any other. I can acknowledge what I want, but once I do, what I have undertaken is not up to me.

Objection 12: I have done nothing to justify the theory of inconsistent concepts I present over any of its competitors. Why isn't some other theory of inconsistent concepts preferable?

Reply 12: See Appendix E, where I compare and contrast it with several other theories.

Objection 13: Consider a language that has a truth predicate that expresses our inconsistent concept of truth. Presumably, classical logic is appropriate for the fragment of the language that does not include this truth predicate. Thus, there are multiple logics for the language as a whole—classical logic and the partial 6-component aletheic logic. It seems that, given a bunch of sentences and arguments from this language, the choice of which logic to use is determined by whether they contain the truth predicate. That is, the aletheic logic overrules classical logic. If an argument does not contain any sentences with truth predicates, then the argument should be evaluated by classical logic, but if it contains even a single sentence with a truth predicate, then one should use the aletheic logic to evaluate it. The problem is that one can find an argument that has no truth predicates and is deductively valid by classical logic (e.g., it is a disjunctive syllogism), but by adding a new premise containing a truth predicate, it becomes an argument that should be evaluated by the aletheic logic and is invalid on this logic. Thus, one looses the monotonicity of deductively valid arguments—one can always turn a valid argument into an invalid one by adding another premise. Moreover, the same considerations hold for the pragmatic theory and the semantic theory. One can find two contexts, one in which a truth predicate is present and the other in which it is not, such that a sentence without a truth predicate is assertible in one but not in the other, and that sentence has one meaning in the first context and a different meaning in the other.

Reply 13: I agree. There are plenty of issues raised by this objection and I cannot address all of them here. If we admit the possibility of defective concepts and we have different logics for them (as we should), then we are going to run into this issue. Whether one counts an argument as valid depends on the inference rules one accepts. If one accepts different inference rules for different topics or contexts, then one will have to face the fact that monotonicity is not a good indicator of deductive validity. On an inferential role account of meaning, the meaning of any given sentence is going to depend on the auxiliary premises available. Thus, people with different beliefs will associate different meanings with a single sentence. The objection merely points out that people that accept different inference rules will associate different meanings to a single sentence, even if they have all the same beliefs. The result does not seem so counterintuitive once it is understood in this way.

7.7 Conclusion

In this chapter, I have applied the theory of inconsistent concepts developed in Chapters Four, Five, and Six to truth to derive a theory of truth, a logic for truth, a pragmatic theory for truth, and a semantic theory for truth. The theory of truth does not give rise to revenge paradoxes or self-refutation problems and the semantic theory for truth satisfies the strong internalizability requirement laid down in Chapters One and Two.

APPENDIX A

FRAGMENTARY THEORIES OF TRUTH

A.1 INTRODUCTION

Some theories of truth explain natural language truth predicates in terms of a group of restricted truth predicates; these restricted truth predicates have extensions that are proper subsets of the extension of 'true'. I call these *fragmentary theories of truth*. Several types of deflationism and the vast majority of approaches to the liar paradox are committed to fragmentary theories of truth. The fragmentary theories of truth I consider are *descriptive* in the sense that they purport to describe the way we actually use truth predicates of natural languages.[1] As such, these theories should be consistent with our intuitions about which uses of truth predicates are proper. In particular, they should be consistent with our intuitions about which assertions of sentences that contain truth predicates are warranted. I argue that many descriptive fragmentary theories of truth imply that some clearly warranted assertions are unwarranted. It follows that these theories are unacceptable.

One way of classifying fragmentary theories of truth is based on the type of restricted truth predicates to which the theory appeals. On this classification scheme, I discuss two types

[1] I use the term 'theory of truth' in a loose way such that a theory of truth describes some aspect of a truth predicate.

of fragmentary theories of truth: (i) those that appeal to language-specific truth predicates (e.g., ‘true-in-English’), and (ii) those that appeal to hierarchies of restricted truth predicates. A second way of classifying fragmentary theories of truth is based on the relation between a natural language truth predicate and the restricted truth predicates to which the theory appeals. I focus primarily on fragmentary theories of truth that treat truth predicates of natural languages as ambiguous; these theories imply that natural language truth predicates take on the meaning of one of the relevant restricted truth predicates on each occasion of use. I also address fragmentary theories of truth that appeal to translation to explain the relation between a natural language truth predicate and the restricted truth predicates in question.

I argue that each of the fragmentary theories of truth I consider is inconsistent with our intuitions about which assertions of truth attributions are warranted (a *truth attribution* is a sentence of the form: p is true). In each case, the theory in question implies that a clearly warranted assertion is unwarranted. The arguments I use for each criticism are instances of a single argument scheme, which I call the *warranted assertibility argument*. Although the warranted assertibility argument casts doubt on a wide range of fragmentary theories of truth, it does not show that all such theories are unacceptable. My objective is to argue for a condition on fragmentary theories of truth: such theories should respect our intuitions about which assertions of sentences that contain truth predicates are warranted. I show that this condition is difficult to meet without betraying the underlying motivations for accepting fragmentary theories of truth.

The appendix is divided into six sections. In the first section, I discuss and motivate fragmentary theories of truth. To guide my presentation, I offer several claims about assertion, warranted assertibility, and ambiguity in section two. The next three sections contain criticisms of fragmentary theories of truth. In section three, I illustrate the warranted assertibility argument by applying it to Alfred Tarski's theory of truth. In section four, I use the argument to criticize those fragmentary theories of truth that explain natural language truth predicates in terms of language-specific truth predicates. In section five, I apply the argument to fragmentary theories of truth that appeal to hierarchies of restricted truth predicates; Hartry Field's recent theory of truth and indeterminacy (which appeals to a transfinite hierarchy of determinate truth predicates) serves as my example. Finally, in section six, I address several objections.

A.2 Fragmentary Theories of Truth

My discussion of the motivations for fragmentary theories of truth is organized by the type of restricted truth predicates to which the theories appeal. Some fragmentary theories of truth explain natural language truth predicates in terms of language-specific truth predicates. I call this explanatory strategy the *language-specific approach* (the *LS approach*).[2] A *language-specific truth predicate* (an *LS truth predicate*) is satisfied only by true sentences of a particular language. For example, 'true-in-English' is a language-specific truth predicate: 'p is true-in-English' is true if p is a true sentence of English, and it is false if p is either a false sentence of English, or a sentence of some other language.[3]

The motivation for the LS approach originates from at least two sources: deflationism and investigations into the logic of truth. *Deflationists* reject explanations of truth in terms of a substantive property or relation (e.g., correspondence with reality, coherence with a body of

[2] I use the term 'approach' as a synonym of 'doctrine' or 'theory'.

[3] I follow most of the theorists I discuss in assuming that sentences (or sentence tokens) are the primary bearers of truth. One can define an LS truth predicate so that attributions containing it are truth-value gaps when the target of the attribution is a sentence of another language. Although I ignore this possibility in the text, my argument could be altered to accommodate it.

belief, utility, etc.). One group of deflationists, *disquotationalists*, favor theories of truth based on the T-sentences, which have the form: ⟨p⟩ is true if and only if p.[4] On disquotational theories of truth, the meaning of a truth predicate is entirely determined by a set of T-sentences (or a T-sentence schema). These truth predicates are defined only for the sentences of a single language; hence, they are LS truth predicates.[5]

The second motivation for the LS approach comes from investigations into the logic of truth. Dating back to Tarski, there is a convention in logic of considering only LS truth predicates. Almost everyone who works on the logic of languages that contain truth predicates follows suit.[6] One explanation for this unanimity is that it is more difficult to provide a logic for an unrestricted truth predicate than for an LS truth predicate; to construct a logic for an unrestricted truth predicate, one must consider its behavior when applied to sentences of every language.

Another reason for focusing exclusively on the logic of LS truth predicates is that it is easier to avoid the liar paradox and its relatives for an LS truth predicate. The liar paradox pertains to *liar sentences*, like the following:

[4] '⟨' and '⟩' are angle quotes; 'p' serves as a sentential variable that can be replaced by a sentence, and '⟨p⟩' is the quote-name of such a sentence. See McGee (1991, 2000) for this usage (McGee uses different symbols). I also use 'p' as a logical constant (e.g.: p is true). Note that these uses are distinct: an occurrence of 'p' cannot be both a sentential variable and a constant. Corner quotes, '⌈' and '⌉', are used in conjunction with constants. For example, if 'p' and 'q' are names of sentences, then '⌈p ∧ q⌉' is the name of the sentence that results from placing sentence p and sentence q on opposite sides of '∧'.

[5] For deflationist theories of this sort, see Leeds (1978, 1995, 1997), Williams (1986, 1999, 2002), Field (1986) (in which they are discussed but not endorsed), Resnik (1990), Quine (1992), McGee (1993), Field (1994a, 1994b), Weir (1996), Halbach (1999, 2000, 2002), and Burgess (2002).

[6] A *logic* is a theory that specifies which arguments in a certain collection are valid. For more examples of logics for truth that address only LS truth predicates, see van Fraassen (1968), Parsons (1974), Kripke (1975), Burge (1979a), Herzberger (1982), Skyrms (1982), Feferman (1982), Yablo (1985), Reinhardt (1986), Priest (1987), McGee (1991), Gaifman (1992), Simmons (1993), Gupta and Belnap (1993), McDonald (2000), Field (2003a, 2003b), and Maudlin (2004). The sentence in the text on which this footnote is a comment contains 'almost' because I am not familiar with every logic for truth, not because I am aware of one that deals with an unrestricted truth predicate for sentences. Barwise and Etchemendy (1987) and Glanzberg (2004) formulate their theories in terms of propositions instead of sentences. Gupta and Belnap (1993: 265-6) speculate on how to define 'true-in-L' where 'L' is a variable that ranges over first-order languages, but that is still not an unrestricted truth predicate for sentences.

(Λ) Λ is false.

The paradox is that an intuitively plausible argument shows that Λ is both true and false.[7] By considering only LS truth predicates, one can ignore liar-like paradoxes that result from inter-linguistic truth attributions (e.g., if p is the German sentence ‘q ist wahr’—which means that q is true—and q is the English sentence ‘p is false’, then both p and q are paradoxical).

Tarski’s theorem on the indefinability of truth provides further motivation for the LS approach. Tarski uses a variant of the liar paradox to prove that for any consistent language L, if L is (i) bivalent (i.e., every sentence of L is either true or false), (ii) mono-aletheic (i.e., no sentence of L is both true and false), (iii) capable of self-reference (i.e., it can quantify over the natural numbers, has a name for zero, and can express addition, multiplication, successor, and identity), and (iv) classical (i.e., it obeys the laws of classical logic), then L cannot contain a truth predicate that is both satisfied by all the true sentences of L and fails to be satisfied by all the false sentences of L.[8] It follows that languages that contain unrestricted truth predicates are either inconsistent or do not satisfy the conditions of Tarski’s theorem. This result suggests that the LS approach offers a promising strategy for solving the liar paradox.

In addition, most approaches to the liar paradox require a distinction between metalanguage and object language.[9] That is, the language in which one of these theories is

[7] The argument is based on the *truth rules* (according to which ⟨⟨p⟩ is true⟩ follows from ⟨p⟩ and vice versa), the *substitution rule* (according to which two names that refer to ⟨p⟩ are intersubstitutable in ⟨⟨p⟩ is true⟩ without changing the truth-value of the sentence), and the inference rules of classical logic. On the one hand, if Λ is true, then ‘Λ is false’ is true (by substitution). If ‘Λ is false’ is true, then Λ is false (by truth rule). Thus, if Λ is true, then Λ is false. On the other hand, if Λ is false, then ‘Λ is false’ is true (by truth rule). If ‘Λ is false’ is true, then Λ is true (by substitution). Thus, if Λ is false, then Λ is true. Therefore, Λ is true if and only if Λ is false. It follows that Λ is both true and false.

[8] Tarski (1933: 247-251). See McGee (1991: ch. 1) and Gupta and Belnap (1993: 49-63) for details on Tarski’s theorem.

[9] Of all the theories cited in footnote 6, only Reinhardt (1986), Priest (1987), McGee (1991), Simmons (1993), Field (2003a, 2003b), and Maudlin (2004) even attempt to dispense with a substantive distinction between object language and metalanguage. In Chapters One and Two, I argue that no semantic theory that requires this distinction is acceptable.

formulated (i.e., the metalanguage) is expressively richer than the languages to which it applies (i.e., the object languages). These approaches are forced to consider only LS truth predicates because permitting an unrestricted truth predicate effectively forfeits the distinction between metalanguage and object language.

To summarize my discussion of the first type of fragmentary theory: both disquotationalism and accounts of the logic of truth predicates provide motivation for the LS approach. The former implies that a truth predicate is defined only for the sentences of a particular language, while the latter benefits from the convenience offered by LS truth predicates and the role LS truth predicates play in approaches to the liar paradox. If disquotationalism or these logics for truth predicates are to be applicable to natural languages, one must be able to explain natural language truth predicates in terms of LS truth predicates.

The second type of fragmentary theory I address explains natural language truth predicates by appeal to a hierarchy of restricted truth predicates. I call this explanatory strategy the *hierarchy approach*. The motivation for the hierarchy approach comes from attempts to solve the liar paradox for languages that are rich in expressive resources. Again, Tarski's theorem is a motivating factor. A consistent language that satisfies the conditions of Tarski's theorem cannot contain even an LS truth predicate that is satisfied by all and only the true sentences of the language. This result suggests that explaining a natural language truth predicate in terms of truth predicates that are restricted to certain sentences of the language in question offers a promising approach to the liar paradox. For example, Tarski shows how to solve the liar paradox for languages rich in expressive resources by treating an LS truth predicate of a language as a hierarchy of truth predicates that are restricted to certain sentences of that language. Saul Kripke shows how to solve the liar paradox without appeal to such a hierarchy;

however, if one applies his theory to languages that have certain expressive resources, new paradoxes that are similar to the liar (called *revenge paradoxes*) emerge. Field shows that one can apply a theory like Kripke's to expressively rich languages if one appeals to a hierarchy of determinate truth predicates to solve the revenge paradoxes.[10] If hierarchy approaches are to be successful, then one must be able to explain a natural language truth predicate in terms of the restricted truth predicates of the hierarchy.

A proponent of a fragmentary theory of truth must answer the question: what is the relation between the restricted truth predicates posited by the theory and a natural language truth predicate? Surprisingly, most proponents of fragmentary theories of truth ignore this crucial issue. One answer is that natural language truth predicates are ambiguous and can have the meaning of any of the restricted truth predicates in question.[11] Another is that a natural language truth predicate is context-dependent—its content changes from context to context, and in any given context, its content is identical to the content of one of the restricted truth predicates.[12] I do not consider context-dependence interpretations of fragmentary theories in this appendix. Instead, I take issue with fragmentary theories of truth that imply that natural language truth predicates are ambiguous (I also consider an LS approach that appeals to translation instead of ambiguity). Before presenting these arguments, I discuss some aspects of assertion, warranted assertibility, and ambiguity.

[10] Tarski (1933), Kripke (1975), and Field (2003a, 2003b). For other proponents of the hierarchy approach, see Fitch (1964), Myhill (1975), Parsons (1974), Burge (1979a), Barwise and Etchemendy (1987), Gupta (1990), Gaifman (1992, 2000), Koons (1992, 2000), Gupta and Belnap (1993), Simmons (1993), Cantini (1995), Williamson (2000b), and Glanzberg (2004, forthcoming). Davidson (forthcoming) tentatively endorses a hierarchy approach to the liar paradox. For criticisms of the hierarchy approach, see Reinhardt (1986), Priest (1987), McGee (1991, 1997), and Simmons (1993). For a defense of the use of hierarchies in approaches to the liar paradox, see Glanzberg (2005).

[11] Parsons (1974) interprets Tarski (1933) in this way; see also Williamson (2000b).

[12] See Parsons (1974), Burge (1979a), Barwise and Etchemendy (1987), Gaifman (1992, 2000), Koons (1992, 2000), Simmons (1993), and Glanzberg (2004). In Appendix B, I argue that approaches to the liar paradox that imply that truth predicates display context-dependence are unacceptable. Although that argument is unrelated to the warranted assertibility argument, it seems to me that a version of the warranted assertibility argument could be used to criticize context-dependence versions of fragmentary theories as well.

A.3 Assertion, Warranted Assertibility, and Ambiguity

In this section, I present several claims about assertion, warranted assertibility, and ambiguity on which I rely in later sections. First, I need an account of the relation between asserting a sentence ⟨p⟩ and asserting that p. I assume that if S asserts ⟨p⟩ in context C and ⟨p⟩ means that q in context C, then S asserts that q in context C. That is a rough characterization, but it will suffice for my purposes.[13]

Second, I require a rudimentary account of warranted assertibility. Philosophers apply 'warrant' to both beliefs and assertions, but the sense in which an assertion is warranted is different from the sense in which a belief is warranted. Beliefs are typically assumed to be attitudes toward propositions, while assertions are actions. The difference between the two is reflected in the areas of philosophy in which they are studied: warranted beliefs are a topic of epistemology, while warranted assertions are usually studied in the philosophy of language or action theory. Roughly, S's *belief that p* is warranted if S has a good reason for believing that p. I treat 'good reason' in a loose way so that the fact that S has a good reason for believing that p can be founded on causal connections, reliability, or other factors of which S might be unaware. Likewise, S's *assertion that p* is warranted if S has a good reason for asserting that p. Again, 'good reason' should be read in a loose way. I assume that a good reason for asserting that p always involves believing that p. That is, I assume that if S asserts that p and S's assertion is

[13] Soames' account of sentences, propositions, and assertions is more precise; see Soames (2002: 105-106).

warranted, then S believes that p.[14] Most theories of warranted assertibility share this assumption.[15]

The warranted assertibility argument appeals only to the belief component of warranted assertibility. In each version of the argument, I present an example of an assertion that is clearly warranted, and I argue that the fragmentary theory of truth in question implies that the asserter does not believe the proposition asserted. Consequently, I could have argued that certain fragmentary theories of truth imply that some sincere assertions are insincere (where an assertion that p is *sincere* if only if the asserter believes that p). I chose to make my point in terms of warranted assertibility because we have firm intuitions about which assertions of truth attributions are warranted, and it is clearly unacceptable if a descriptive theory of truth is inconsistent with these intuitions.

Turning now to ambiguity: a linguistic expression is *ambiguous* if and only if it has two or more determinate, independent meanings. The standard example in English is 'bank', which can mean *effluvial embankment* or *financial institution* (of course, it has other meanings as well).[16] It is important to distinguish ambiguity from confusion, vagueness, and context-dependence. An expression is *confused* if and only if, in employing it, one is committed to applying it to two or more distinct entities without properly distinguishing between them; in other words, the employer of a confused expression thinks that two or more distinct entities are identical. An example is 'mass' as it was used in Newtonian mechanics. Today we know that physical objects have relativistic mass and proper mass (which are different physical properties),

[14] I assume that the relevant notion is *de dicto* belief; I discuss alternatives in section six. I ignore issues associated with expressivism.

[15] For proponents of the knowledge theory of assertion, see Unger (1975: ch. 6), Williamson (1996, 2000a: ch. 11), DeRose (2002), and Hawthorne (2004: 21-24, 85-91). See Wright (1992) and Price (1998) for alternatives. Kripke is a supporter of the link between assertion and belief; he endorses what he calls the *disquotation principle*: if a normal English speaker, on reflection, sincerely asserts ⟨p⟩, then he believes that p; see Kripke (1979).

[16] I am assuming a standard distinction between expression types and tokens. Strictly speaking, expression types are ambiguous.

but before the advent of relativistic physics, people who employed ‘mass’ used it in an attempt to designate what they mistakenly took to be a single property. An expression is v*ague* if and only if it has borderline cases (i.e., it neither definitely applies nor definitely fails to apply to some entities). An expression is *context-dependent* if and only if its content depends on the context in which it is used. It is common to distinguish the meaning of a context-dependent expression, which is constant, from its content, which varies. According to my usage, a linguistic expression can be ambiguous without being confused, vague, or context-dependent.[17]

A.4 THE TARSKIAN APPROACH

A Tarskian theory of truth for a natural language is fragmentary in two senses: it is both an LS approach and a hierarchy approach. That is, the truth predicate of the natural language is explained in terms of a set of language-specific truth predicates, and each of these is explained in terms of a hierarchy of more restricted truth predicates. In this section, I concentrate on the hierarchy aspect of Tarski’s theory.

Tarski proposes a theory of truth for certain formal languages that do not contain their own truth predicates. Of course, natural languages seem to contain their own truth predicates (e.g., English speakers apply ‘true’ to sentences of English). To apply Tarski’s theory of truth to a language like English, one constructs a hierarchy of truth predicates that are restricted to certain sentences of the language. For example, a Tarskian theory of truth for English appeals to a hierarchy of truth predicates (true$_0$, true$_1$, etc.), which are restricted to certain sentences of English depending on the sentences’ levels. A sentence of English is level 0 if and only if it

[17] See Atlas (1989) on ambiguity, Field (1973) and Camp (2002) on confusion, Williamson (1994) on vagueness, and Kaplan (1989) on context-dependence.

contains no truth predicate. Roughly, a sentence of English is level n only if it attributes truth to sentences whose levels are less than n, and one of the sentences to which it attributes truth is level n – 1. The predicate $\ulcorner$true$_n\urcorner$ is satisfied by a sentence only if the sentence's level is less than or equal to n. For example, if p is a true English sentence of level 0, then 'p is true$_0$' is true; if q is a true English sentence of level 3, then 'q is true$_2$' is false.[18] I call true$_0$, true$_1$, etc., *Tarskian truth predicates*.

One might object to Tarski's theory of truth on the grounds that English contains a single truth predicate, 'true', not an infinite number of Tarskian truth predicates. A proponent of the Tarskian approach might reply by stipulating that 'true' of English is ambiguous—it takes on the meaning of one of the Tarskian truth predicates on each occasion of use. I call this the *ambiguity Tarskian approach* (the *AT approach*).[19]

The variant of the warranted assertibility argument that undermines the AT approach is inspired by one of Kripke's criticisms of Tarski. The following is a portion of Kripke's remarks on the AT approach:

> If someone makes such an utterance as (1) [i.e., 'Most (i.e., a majority) of Nixon's assertions about Watergate are false'][20], he does *not* attach a subscript, explicit or implicit, to his utterance of 'false', which determines the "level of language" on which he speaks. An implicit subscript would cause no trouble if we were sure of the "level" of *Nixon's* utterances; we could then cover them all, in the utterance of (1) or even of the stronger
>
> (4) All of Nixon's utterances about Watergate are false.
>
> simply by choosing a subscript higher than the levels of any involved in Nixon's Wategate-related utterances. Ordinarily, however, a speaker *has no way of knowing the "levels" of Nixon's relevant utterances*. … If the speaker is forced to assign a "level" to (4) in advance [or to the word 'false' in (4)][21], he may be

[18] Tarski (1933). This approach avoids the liar paradox because Λ is a sentence of level n that attributes truth$_n$ to a sentence of level n (itself); hence, it is false. See Church (1976), Halbach (1997), and Soames (1999) for more details on the Tarskian approach.

[19] See Kripke (1975: 695) for discussion of this interpretation of Tarski's theory; Kripke attributes this interpretation to Parsons (1974). Other interpretations of the Tarskian hierarchy are possible as well. See Burge (1979a, 1982a, 1982b) for a reading on which a truth predicate of a natural language is an indexical.

[20] I have added the bracketed text to indicate Kripke's sentence (1).

[21] This bracketed text is in the original.

> unsure how high a level to choose; if, in ignorance of the "level" of Nixon's utterances, he chooses too low, his utterance of (4) will fail of its purpose, (Kripke 1975: 695-696).

Although Kripke does not phrase his criticism in terms of warranted assertibility or provide an argument for his conclusion, he does mention that in ordinary situations English speakers would have trouble using an ambiguous truth predicate of this sort. In addition to explicitly arguing against the AT approach, I propose the account of fragmentary theories of truth, and I use the warranted assertibility argument to criticize a wide range of fragmentary theories of truth (including the LS approach, which Kripke endorses). I certainly do not want to attribute these proposals to Kripke. However, it is unclear to me whether the rendition of the warranted assertibility argument I use against the AT approach is what Kripke had in mind as a criticism of Tarski. If I may parody Kripke: probably many of my formulations and recastings of the argument are done in a way Kripke would not himself approve. So the present *section* should be thought of as expounding neither 'Kripke's' argument nor 'Scharp's': rather Kripke's argument as it struck Scharp, as it presented a problem for him.[22]

The following situation sets up the first instance of the warranted assertibility argument. Ned and Maude are at a bar, on Tuesday, having a conversation in English. Maude is a distinguished expert on ring-tailed lemurs, and Ned is aware of this fact. Maude tells Ned that on Monday she was at a talk given in English by Helen, another expert on ring-tailed lemurs. Maude informs Ned that Helen argued for a certain thesis, but Maude does not tell Ned what the thesis is. Maude simply refers to it as *Helen's thesis*. Maude remarks that Helen's thesis implies that a theory Maude recently published is false, and she tells Ned that she now agrees with

[22] The last two sentences are, of course, a parody of a famous passage in Kripke's monograph on Wittgenstein's private language argument: "Probably many of my formulations and recastings of the argument are done in a way Wittgenstein would not himself approve. So the present paper should be thought of as expounding neither 'Wittgenstein's' argument nor 'Kripke's': rather Wittgenstein's argument as it struck Kripke, as it presented a problem for him," (Kripke 1982: 5).

Helen.[23] Later that morning, Ned bumps into Tim at the library. Tim is writing a paper on ring-tailed lemurs, and he informs Ned that he is planning to rely on Maude's recently published theory. Ned tells Tim that Maude's theory is false. Tim knows that Ned is usually sincere and trustworthy, but that Ned does not know much about the literature on ring-tailed lemurs; accordingly, Tim challenges Ned on his assertion. Ned responds by asserting 'if Helen's thesis is true, then Maude's theory is false' and 'Helen's thesis is true'. Ned, of course, explains to Tim that Maude told him of these facts. After hearing this, Tim scurries off to the bar to find Maude so that he can find out what Helen's thesis is. (I refer to the conversation between Tim and Ned as *context C*.)

Ned believes that Helen's thesis is true, and Ned's assertion of 'Helen's thesis is true' is warranted. However, Ned does not have beliefs about which sentence Helen's thesis is, the level to which it belongs, or which Tarskian truth predicate it satisfies.[24] This example shows that, given the established usage of 'true' among English users, it is possible that (i) S asserts that a sentence p is true, (ii) S's assertion is warranted, and (iii) it is not the case that there is some positive integer *i* such that S believes that p is true_i.[25]

The AT approach implies that Ned's assertion is unwarranted, as the following argument shows:

[23] I assume that Maude is right about the truth of Helen's thesis and its consequence.

[24] One might object that Ned should believe that Helen's thesis is level 0. In response, I point out that Helen's thesis could contain a truth predicate (e.g., it could be 'if Maude's theory is true, then ring-tailed lemurs are not primates').

[25] One might object to my formulation on Quinean grounds that it is illegitimate to quantify into belief contexts because of their opacity; see Quine (1956). My reply: the quantification into belief contexts in my example is optional. I formulate it in this way for its convenience and precision. Quine distinguishes between *notional* belief and *relational* belief (some philosophers call these belief *de dicto* and belief *de re*), and rejects quantification into belief contexts as a way of explaining relational belief. This distinction is based on quantification into the object position (e.g., for all x, if x is x monkey, Ned believes that x is cute). In my formulation, there is quantification into the subscript of the predicate position (e.g., for some x, if x is a positive integer, then it is not the case that Ned believes that q is true_x). Moreover, several philosophers have defended the legitimacy of quantifying into belief contexts; see Hintikka (1962), Kaplan (1969, 1986), Forbes (1996), and Santambrogio (2002). For criticism of quantifying into belief contexts, see Tienson (1987). For more on the issue of *de re* and *de dicto* belief, see Burge (1977), McDowell (1984), and Brandom (1994: 495-573).

(AT1) Ned asserts 'Helen's thesis is true' (in context C).

(AT2) On the AT approach, for some positive integer *i*, 'Helen's thesis is true' means that Helen's thesis is true$_i$ (in context C).

(AT3) On the AT approach, for some positive integer *i*, Ned asserts that Helen's thesis is true$_i$ (in context C).

(AT4) For every positive integer *i*, if Ned asserts that Helen's thesis is true$_i$ and Ned's assertion is warranted, then Ned believes that Helen's thesis is true$_i$.

(AT5) It is not the case that for some positive integer *i*, Ned believes that Helen's thesis is true$_i$.

∴ (AT6) On the AT approach, Ned's assertion is unwarranted.

I call this the *warranted assertibility argument.*[26]

(AT1) is true by stipulation. (AT2) follows from the definition of the AT approach. (AT3) follows from (AT2) and the assumption about assertion I proposed in section two. (AT4) follows from the assumption about warranted assertibility I proposed in section two. That leaves (AT5) open for the AT theorist to reject.

If the AT theorist rejects (AT5), then she suggests that for some positive integer *i*, Ned believes that Helen's thesis is true$_i$. Thus, for some positive integer *i*, Ned asserts that Helen's thesis is true$_i$, and he believes that Helen's thesis is true$_i$; hence, his assertion is warranted. My reply to this suggestion is that it attributes to Ned a belief he does not have; in the example, Ned believes that Helen's thesis is true, not that Helen's thesis is true$_i$ (for some positive integer *i*). Ned does not know what Helen's thesis is, much less what level it has or which Tarskian truth

[26] The warranted assertibility argument highlights one of the problems with treating univocal expressions as ambiguous. As such, it complements the following famous remark of Kripke's: "it is very much the lazy man's approach in philosophy to posit ambiguities when in trouble. If we face a putative counterexample to our favorite philosophical thesis, it is always open to us to protest that some key term is being used in a special sense, different from its use in the thesis. We may be right, but the ease of the move should counsel a policy of caution: Do not posit an ambiguity unless you are really forced to, unless there are really compelling theoretical or intuitive grounds to suppose that an ambiguity is really present," (Kripke 1977: 19); Anscombe expresses a similar sentiment in the following passage: "where we are tempted to speak of 'different senses' of a word which is clearly not equivocal, we may infer that we are pretty much in the dark about the concept it represents," (Anscombe 1957: 1).

predicate is appropriate for it. Thus, this version of the AT approach implies that Ned has a belief he does not have.

A proponent of the AT approach can, of course, stipulate that for some positive integer *i*, Ned's belief that Helen's thesis is true is identical to the belief that Helen's thesis is true_i. However, this account attributes the wrong content to Ned's belief. In particular, this theory implies that if Ned believes that Helen's thesis satisfies some Tarskian truth predicate or other, then he believes that Helen's thesis satisfies a particular Tarskian truth predicate. This result is intolerable. Hence, rejecting (AT5) is unacceptable. (The AT theorist could insist that, despite the evidence to the contrary, for some positive integer *i*, Ned does believe Helen's thesis is true_{i}; I consider this option in section six.)

Instead of rejecting (AT5), a proponent of Tarski's theory might revise the AT approach by permitting 'true' to have meanings other than those of the Tarskian truth predicates. One choice is to stipulate that on an occasion of use, 'true' can have either the meaning of any of the Tarskian truth predicates or the meaning of a generic Tarskian truth predicate (whose extension is the union of the extensions of the Tarskian truth predicates, and whose anti-extension is the intersection of their anti-extensions); one could achieve the same results by permitting quantification into the subscripts of the Tarskian truth predicates. A proponent of the revised AT approach can then claim that Ned asserts that Helen's thesis is generically true, and Ned believes that Helen's thesis is generically true; hence, his assertion is warranted.

In response, I argue that the AT approach cannot admit a generic Tarskian truth predicate (or quantification in to the subscripts of the Tarskian truth predicates) because this addition would introduce a relative of the liar paradox. With a generic truth predicate available, one can construct the following sentence:

(Γ) Γ is not generically true.

Γ means that it is not the case that for some positive integer *i*, Γ is true$_i$. According to the assumptions of Tarski's theory, Γ is both generically true and not generically true. Assume that Γ is generically true. If so, then for some *i*, Γ is true$_i$. If for some *i*, Γ is true$_i$, then for some *i*, 'Γ is not generically true' is true$_i$. If for some *i*, 'Γ is not generically true' is true$_i$, then Γ is not generically true. Hence, if Γ is generically true, then Γ is not generically true. Assume that Γ is not generically true. If so, then for some *i*, 'Γ is not generically true' is true$_i$. If for some *i*, 'Γ is not generically true' is true$_i$, then for some *i*, Γ is true$_i$. If for some *i*, Γ is true$_i$, then Γ is generically true. Hence, if Γ is not generically true, then Γ is generically true. Therefore, Γ is generically true if and only if Γ is not generically true. It follows that Γ is both generically true and not generically true. Therefore, introducing a generic Tarskian truth predicate would undermine the primary motivation for the AT approach.

A.5 Language-Specific Approaches

In the previous section, I presented the warranted assertibility argument, and I used it to criticize the AT approach. In this section, I apply the warranted assertibility argument to certain fragmentary theories of truth that exemplify the language-specific approach (the LS approach). On the LS approach, one can explain natural language truth predicates in terms of language-specific truth predicates (LS truth predicates). In the next section, I address the hierarchy approach.

An objection to the LS approach is that natural language truth predicates are not language-specific. For example, one can attribute truth to a German sentence by using the truth predicate of English (e.g., ' ' Schnee ist weiss' is true' is a true sentence of English). A proponent of the LS approach can respond by claiming that 'true' in English is ambiguous and can have the meaning of any of the LS truth predicates. I call such a truth predicate an *ambiguous language-specific truth predicate* (an *ALS truth predicate*), and I call the species of the LS approach on which natural language truth predicates are ALS truth predicates, the *ambiguity language-specific approach* (the *ALS approach*). I now turn to a version of the warranted assertibility argument that undermines the ALS approach.

Recall the example with Ned, Maude, Tim, and Helen. In this version, Maude, the expert on ring-tailed lemurs, is fluent in many languages; Ned is aware of this fact, but he does not know which particular languages Maude comprehends (other than English). Maude tells Ned about Helen's thesis without telling him what the thesis is or the language Helen was speaking when she asserted it. Ned has the same interaction with Tim that he had in the first version. In particular, Ned asserts 'Helen's thesis is true', and this assertion is warranted. It is not the case that Ned has a belief about either the languages in which Helen's talk was given or the LS truth predicate Helen's thesis satisfies. This example shows that, given the established usage of 'true' among English users, it is possible that (i) S asserts that a sentence p is true, (ii) S's assertion is warranted, and (iii) it is not the case that there is some language *l* such that S believes that p is true-in-*l*.

The ALS approach implies that Ned's assertion is unwarranted, as the following version of the warranted assertibility argument shows:

(ALS1) Ned asserts 'Helen's thesis is true' (in context C).

(ALS2) On the ALS approach, for some language *l*, 'Helen's thesis is true' means that Helen's thesis is true-in-*l* (in context C).

(ALS3) On the ALS approach, for some language *l*, Ned asserts that Helen's thesis is true-in-*l* (in context C).

(ALS4) For every language *l*, if Ned asserts that Helen's thesis is true-in-*l* and Ned's assertion is warranted, then Ned believes that Helen's thesis is true-in-*l*.

(ALS5) It is not the case that for some language *l*, Ned believes that Helen's thesis is true-in-*l*.

∴ (ALS6) On the ALS approach, Ned's assertion is unwarranted.

The justifications for the premises of this argument are analogous to those given in section three. In an attempt to avoid (ALS5), an advocate of the LS approach might revise the ALS approach by introducing a generic language-specific truth predicate (whose extension is the union of the extensions of the LS truth predicates, and whose anti-extension is the intersection of their anti-extensions); one could achieve the same results by allowing quantification into the language position of 'true-in-L'. I see no reason in principle why this cannot be done. However, I am not aware of an advocate of the LS approach who endorses this strategy, and there are good reasons for avoiding it. First, one would have to either give an account of quantification over languages or define an LS truth predicate for every language; either project would require a substantial theory of language. That is not impossible, but a sufficiently precise and plausible account of language would be difficult to produce. Moreover, it is unclear whether a deflationist who advocates the LS approach could appeal to a notion of truth conditions that would be sufficient to individuate languages in the right way.[27] In addition, appeal to a generic LS truth predicate would undermine another motivation for the LS approach: namely, if one added a generic LS truth predicate or allowed quantification into the language position of 'true-in-L', then the LS approach would be susceptible to the liar paradox (see section one).

[27] David (1994: 158-166) makes a similar point.

Some deflationists who advocate the LS approach suggest that natural language truth predicates should be understood as translational language-specific truth predicates.[28] A *translational language-specific truth predicate* (a *TLS truth predicate*) for a language L is a predicate that is synonymous with 'translatable into a sentence of L that is true-in-L'. For example, a proponent of treating natural language truth predicates as TLS truth predicates claims that the English sentence ' ' Schnee ist weiss' is true' means that 'Schnee ist weiss' is translatable into an English sentence that is true-in-English. I call this species of the LS approach the *translational language-specific approach* (the *TLS approach*).[29]

There are at least two options for the way the TLS approach interprets an English sentence like ' ' Schnee ist weiss' is true': the *quantificational version*, which treats this sentence as '(∃x)(x is a sentence of English and x is a translation of 'Schnee ist weiss' and x is true-in-English)', and the *constant version*, which treats it as 'p is a sentence of English and p is a translation of 'Schnee ist weiss' and p is true-in-English', where 'p' is a constant. When interpreting multiple-target truth attributions (e.g., 'all the sentences Carl asserted yesterday are true'), the quantificational version of the TLS approach is the only acceptable option.[30] Thus, one might as well endorse it in general.

[28] The move is familiar in the face of other criticisms leveled against LS approaches; for such criticisms, see David (1989, 1994), Richard (1996), Soames (1997), Brendel (2000), Horwich (2001), Azzouni (2001), Künne (2002), and Shapiro (2003). These philosophers all address deflationists who advocate the LS approach. Some philosophers who work on the logic of truth have criticized Tarski's commitment to the LS approach; see Field (1972), Dummett (1978: introduction), and Putnam (1985). See also Davidson (1990, forthcoming) for discussion of this issue.

[29] On the TLS approach, see Field (1986) (in which it is discussed but not endorsed), McGee (1993), Field (1994a) (in which it is endorsed as an option), Leeds (1995, 1997), and Williams (1999, 2002). For deflationist alternatives to LS truth predicates, see Field (1994a), Lance (1996), Azzouni (2001), Horwich (2001), and Brandom (2002).

[30] The quantificational version can render this claim as: 'for all x, if x is a sentence Carl asserted yesterday, then for some y, y is a sentence of English and y is a translation of x and y is true-in-English'. How should the constant version treat this sentence? Perhaps 'for all x, if x is a sentence Carl asserted yesterday, then p is a sentence of English and p is a translation of x and p is true-in-English'? This cannot be right because it implies that p is a translation of all the sentences Carl asserted. Another option might be: 'for all x, if x is a sentence Carl asserted yesterday, then p and q are sentences of English, and either p or q is a translation of x, and p and q are true-in-English'. This suggestion works only if Carl asserted at most two sentences on the day in question. A supporter of the constant version might suggest that the logical form of the truth attribution depends on the number of its targets,

The TLS approach is plausible only for accounts of language and accounts of translation on which all languages are intertranslatable. Otherwise, it faces an obvious criticism. Pick a true sentence p of a language L that is not translatable into English. The sentence ‘p is true’ is a true sentence of English, but the TLS approach implies that it is false (on the TLS approach, ‘p is true’ means that for some x, x is a sentence of English and x is a translation of p and x is true-in-English—but, by stipulation, there is no such sentence of English). Thus, I assume that, given the notions of language and translation employed by the TLS theorist, all languages are intertranslatable.

This concession does not save the TLS approach from the warranted assertibility argument. Again, the example is a story with Ned, Maude, Helen, and Tim. In this version, Maude tells Ned about Helen’s thesis, but she does not tell him which language Helen was speaking when Helen asserted it. Maude tells Ned that Helen’s thesis cannot be translated into English because it involves technical jargon that currently belongs only to the language Helen was speaking when she asserted it. When Ned meets up with Tim, Ned asserts ‘Helen’s thesis is true’, and his assertion is warranted. Furthermore, Ned believes that Helen’s thesis is not translatable into English; consequently, Ned believes that there is no sentence of English that is both a translation of Helen’s thesis and true-in-English.

The following version of the warranted assertibility argument shows that, on the TLS approach, Ned’s assertion is unwarranted:

(TLS1) Ned asserts ‘Helen’s thesis is true’.

(TLS2) On the TLS approach, ‘Helen’s thesis is true’ means that for some x, x is a sentence of English and x is a translation of Helen’s thesis and x is true-in-English.

but this hardly seems plausible. Moreover, it abandons the view that ‘true’ means ‘translatable into a sentence of English that is true-in-English’.

(TLS3) On the TLS approach, Ned asserts that for some x, x is a sentence of English and x is a translation of Helen's thesis and x is true-in-English.

(TLS4) If Ned asserts that for some x, x is a sentence of English and x is a translation of Helen's thesis and x is true-in-English, and Ned's assertion is warranted, then Ned believes that for some x, x is a sentence of English and x is a translation of Helen's thesis and x is true-in-English.

(TLS5) It is not the case that Ned believes that for some x, x is a sentence of English and x is a translation of Helen's thesis and x is true-in-English.

∴ (TLS6) On the TLS approach, Ned's assertion is unwarranted.

Again, the justifications for the premises are analogous to the ones given in section three. Unlike both the AT and ALS approaches, there is no generic truth predicate to consider for the TLS approach.

A.6 Hierarchy Approaches

In the previous section, I applied the warranted assertibility argument to certain LS approaches. In this section, I apply it to another type of fragmentary theory: the hierarchy approach. A hierarchy approach explains a truth predicate of a natural language in terms of a hierarchy of restricted truth predicates. Because the principal motivation for the hierarchy approach arises from attempts to solve the liar paradox, the truth predicates to which the theory appeals are almost always used to classify paradoxical sentences. My example of the hierarchy approach is Field's theory of truth and indeterminacy.

In a series of recent papers, Field introduces an impressive account of partially defined expressions. From this account he derives a theory of truth and indeterminacy, a theory of vagueness, and a theory of properties.[31] He uses his account of partially defined expressions to provide a novel, powerful, and unified solution to the liar paradox, Curry's paradox, the sorites paradox, and the property version of Russell's paradox.[32] To accompany his account, he presents a new formulation of deflationism, a new non-classical logic with an intuitive conditional, and a non-standard probability calculus that allows him to explain degrees of belief in propositions that display indeterminacy.[33] There is no question that Field's account of partially defined expressions deserves to be one of the most discussed topics in the philosophy of language for years to come. It should be clear that a discussion of this entire account is beyond the scope of this appendix. Instead, I focus on his theory of truth and indeterminacy, but I provide only the details that are relevant to my criticism.

Field begins with a version of Kripke's theory of truth. Kripke was one of the first to present a formal theory of truth that applies to languages that contain their own truth predicates; he accomplishes this feat by allowing truth-value gaps in the languages he considers. A language displays truth-value gaps if it contains sentences that are in neither the extension nor the anti-extension of 'true'.[34] A problem with Kripke's approach is that the languages he considers lack an intuitive conditional. One of Field's innovations is showing how to add a

[31] See Field (2002, 2003a, 2003b, 2003c, forthcoming a, forthcoming b) for the theory of truth and indeterminacy, Field (2003b, 2003c) for the theory of vagueness, and Field (2003c, 2004) for the theory of properties.

[32] I discussed the liar paradox in section one. Curry's paradox is that, from intuitive assumptions, one can use the sentence 'if this sentence is true, then God exists' to derive that God exists (or any other absurdity). The sorites paradox is that, from intuitive assumptions, one can derive that a person with a full head of hair is bald (the reasoning works for most vague expressions, not just 'bald'). The property version of Russell's paradox is that, from intuitive assumptions, one can derive that the property of non-self-instantiation both instantiates itself and does not instantiate itself. It seems that Field's account could be adapted to provide solutions for the paradoxes of denotation (Richard's paradox, Berry's paradox, and König's paradox) and Grelling's paradox of predication as well, but he does not emphasize this aspect of his work. Field is pessimistic about using it as a solution to the set-theoretic paradoxes; however, he also does not think that a new solution is in order. See Field (2004).

[33] See the previously cited works on his theory of truth for the first two topics and Field (2000, 2001b, 2003b, forthcoming c) for the non-standard probability calculus. One should be aware that Field presents two nonstandard probability calculi, one that is classical and the other nonclassical. He now endorses only the nonclassical version; see Field (2003c: 462).

[34] See Kripke (1975). Kripke considers only LS truth predicates, but I am ignoring this complication in this section.

conditional to these languages that obeys most of the principles we associate with conditionals (e.g., it obeys modus ponens, '$A \rightarrow A$' is a logical truth, etc.).

Field's method for adding a truth predicate and a conditional to a first-order language results in a language that has truth-value gaps and for which the principle of excluded middle (i.e., $\ulcorner p \vee \sim p \urcorner$) fails in general.[35] The conditional behaves just like a material conditional if one assumes the relevant instances of the principle of excluded middle. According to Field, truth displays indeterminacy in the sense that some sentences (e.g., the liar sentence) are in neither the extension nor the anti-extension of 'true'; however, neither 'the liar sentence is not true', nor 'the liar sentence is not false' is a member of the extension of 'true' because the sentence 'the liar sentence is either true or false' is indeterminate. Instead, one can describe the status of the liar sentence by asserting that it is not determinately true and not determinately false (i.e., the liar sentence is indeterminate). One of the most satisfying aspects of Field's theory is that he provides the means for asserting that the liar sentence is indeterminate in the languages he considers. He defines a sentential determinacy operator, 'D', in terms of his conditional:

$$DA =_{df} A \wedge (\top \rightarrow A).$$

Here, '$\top$' is any tautology (e.g., '$A \rightarrow A$' or '$0 = 0$').

That Field provides a characterization of the liar in the languages he considers is a huge advance over Kripke's theory, which implies that the liar is a truth-value gap, but cannot be applied to languages that contain truth-value gap predicates. The problem with applying Kripke's theory to such languages is that one can use a truth-value gap predicate to construct a

[35] Field denies that the languages he considers have truth-value gaps; see Field (2003b: 270). However, Field means something different by 'true-value gap' than most of those who work on truth and the liar paradox. The common usage is that a sentence p of a language L is a truth-value gap if p is a member of neither the extension nor the anti-extension of 'true-in-L'. On Field's usage, a sentence p of L is a truth-value gap if the sentence 'p is not true-in-L and p is not false-in-L' is in the extension of 'true-in-L'. The languages Field considers allow truth-value gaps in the common sense of the term, but do not allow truth-value gaps in the Fieldian sense of the term.

revenge paradox for his theory.[36] Given that Field's theory implies that the liar sentence is indeterminate (not determinately true and not determinately false), the revenge paradox for Field's theory of truth involves the sentence:

(Δ) Δ is either indeterminate or false.

I call Δ the *determinate liar*. Field is ready for it. Although the sentence 'Δ is determinately true or determinately false' is indeterminate, Field's determinacy operator iterates non-trivially so that one can say that Δ is not *determinately* determinately true and not determinately false (determinate determinate falsity is equivalent to determinate falsity, but determinate determinate truth is stronger than determinate truth).[37] Of course, one can formulate a determinate determinate liar, but Field's determinacy operator iterates again so that he can characterize it in the language as well. In fact, Field shows how to define a transfinite hierarchy of determinacy operators (D^{σ}) in terms of which he defines a transfinite hierarchy of determinate truth predicates (D^{σ}true(x)) and a transfinite hierarchy of indeterminacy predicates ($\sim D^{\sigma}$true(x) $\wedge$ ~Dfalse(x)). When I say that Field constructs a transfinite hierarchy of determinacy operators, I mean that, given a system of ordinal notations (roughly, a certain mapping from a set of integers onto a segment of the ordinals), Field shows how to construct an operator $\ulcorner D^{\sigma} \urcorner$ for any ordinal σ in a proper initial segment of the recursive ordinals. There is no maximal recursively related system of ordinal notations; hence, given a system of ordinal notations for a proper initial segment of the

[36] A *revenge paradox* for a theory of truth is a paradox that is similar to the liar, but employs a key expression of the theory in question. Kripke's theory characterizes the liar sentence as gappy; the revenge liar for Kripke's theory involves sentence g: g is either gappy or false. If g is either gappy or false, then 'g is either gappy or false' is true. If 'g is either gappy or false' is true, then g is true. Hence, if g is either gappy or false, then g is true. Likewise, if g is true, then 'g is either gappy or false' is true. If 'g is either gappy or false' is true, then g is either gappy or false. Hence, if g is true, then g is either gappy or false. Consequently, g is true if and only if g is either gappy or false. Therefore, g is both true and either gappy or false (a contradiction—if we assume that a sentence that is either gappy or false is not true).

[37] It might be helpful to see that 'determinately determinately A' is synonymous with '$(A \wedge (0 = 0 \rightarrow A)) \wedge (0 = 0 \rightarrow (A \wedge (0 = 0 \rightarrow A)))$'.

recursive ordinals, one can construct a new system of ordinal notations for a larger proper initial segment of the recursive ordinals. One can construct a new set of determinacy operators based on the new system of ordinal notations such that there exists an operator $\ulcorner D^{\sigma} \urcorner$ in the new set that is not a member of the old set. In less precise language: there are many different ways of constructing the hierarchy of determinacy operators, and there is no "highest" hierarchy of them—given one hierarchy of determinacy operators, one can construct a "higher" one.[38]

In the remainder of this section, I am concerned with showing that a version of the warranted assertibility argument casts doubt on Field's theory. (Because Field advocates a language-specific approach to natural language truth predicates, the criticism presented in the previous section applies to his theory as well.) When applying his theory to a natural language Field faces an objection that is analogous to the one we saw leveled against both the Tarskian approach and the language-specific approach: namely, English contains a single expression 'indeterminate', not a transfinite hierarchy of indeterminacy predicates.[39] A proponent of Field's theory might respond to this criticism with the claim that our word 'indeterminate' in English is ambiguous and can take on the meaning of any of the Fieldian indeterminacy predicates in the hierarchy. However, unlike any of the previously considered approaches, Field's theory permits quantification into the subscripts of the restricted truth predicates to which his theory appeals.

[38] See Field (2003a, 2003b, forthcoming b: fn. 14, fn. 21). I want to thank Hartry Field for conversations on this aspect of his theory in which he pointed out several mistakes in my exposition. See Rogers (1967: 205-213), and Setzer (1999), and Rathjen (1999) on ordinal notations. Those who are familiar with approaches to the liar paradox will no doubt be able to predict the next problem for Field. One can generate a new revenge paradox by quantifying into the subscripts of the indeterminacy predicates (assuming a fixed system of ordinal notations). Consider: 'either this sentence is false or for all ordinals σ, this sentence is indeterminate$_{\sigma}$'. Field responds to this problem by claiming that there is no way to introduce a "maximal" determinacy operator into the language. He also appeals to Tarski's indefinability theorem in an effort to persuade us that we should not expect such an operator to be definable in the language. See Field (2003a, 2003b). Evaluating this response is beyond the scope of this appendix, but my arguments do not depend on whether it is adequate. See Yablo (2003) for criticism of this aspect of Field's theory.

[39] I present the warranted assertibility argument against Field's theory by focusing on his interpretation of 'indeterminate'; I could have considered the implications of his theory for 'true' instead. Because he explains his indeterminacy predicates in terms of truth, this choice makes no real difference.

Thus, a defender of Field's theory can claim that 'indeterminate' is ambiguous and can have *either* the meaning of one of the Fieldian indeterminacy predicates or the meaning of 'indeterminate$_\sigma$ for some ordinal σ'.[40] I call this the *ambiguity Fieldian approach* (the *AF approach*).

Consider once more the example with Ned, Maude, Tim, and Helen. In this version, Maude tells Ned that Helen's thesis is indeterminate; Ned believes her, and he has good reason to believe her. Ned asserts 'Helen's thesis is indeterminate' to Tim. Ned has no beliefs about what Helen's thesis is, what its level of indeterminacy is, or which Fieldian indeterminacy predicate it satisfies.

The following version of the warranted assertibility argument shows that the AF approach implies that Ned's assertion is unwarranted (assume a fixed system of ordinal notations):

(AF1) Ned asserts 'Helen's thesis is indeterminate' (in context C).

(AF2) On the AF approach, either for some ordinal σ, 'Helen's thesis is indeterminate' means that Helen's thesis is indeterminate$_\sigma$ (in context C), or 'Helen's thesis is indeterminate' means that for some ordinal σ, Helen's thesis is indeterminate$_\sigma$ (in context C).

(AF3) On the AF approach, either for some ordinal σ, Ned asserts that Helen's thesis is indeterminate$_\sigma$ (in context C), or Ned asserts that for some ordinal σ, Helen's thesis is indeterminate$_\sigma$ (in context C).

(AF4) For every ordinal σ, if Ned asserts that Helen's thesis is indeterminate$_\sigma$ and Ned's assertion is warranted, then Ned believes that Helen's thesis is indeterminate$_\sigma$; if Ned asserts that for some ordinal σ, Helen's thesis is indeterminate$_\sigma$, and Ned's assertion is warranted, then Ned believes that for some ordinal σ, Helen's thesis is indeterminate$_\sigma$.

[40] As far as I know, Field does not endorse this interpretation. He is careful to say that he is "non-committal" on the issue of whether his determinacy operator "is a fully accurate reflection of 'the' intuitive notion of determinacy," (Field 2003c: 479).

(AF5) It is not the case that for some ordinal σ, Ned believes that Helen's thesis is indeterminate$_{\sigma}$, and it is not the case that Ned believes that for some ordinal σ, Helen's thesis is indeterminate$_{\sigma}$.

∴ (AF6) On the AF approach, Ned's assertion is unwarranted.

The justifications for (AF1) through (AF4) are analogous to those given in section three.

A proponent of Field's theory will almost certainly reject the second conjunct of (AF5). Hence, the AF theorist claims that Ned asserts that for some ordinal σ, Helen's thesis is indeterminate$_{\sigma}$, and Ned believes that for some ordinal σ, Helen's thesis is indeterminate$_{\sigma}$. Thus, on this objection, Ned's assertion is warranted. In response, I point out that this suggestion attributes to Ned a belief he does not have. Ned believes that Helen's thesis is indeterminate. It is not the case that Ned believes that for some ordinal σ, Helen's thesis is indeterminate$_{\sigma}$.

A proponent of the AF approach can, of course, stipulate that Ned's belief that Helen's thesis is true is identical to the belief that for some σ, Helen's thesis is indeterminate$_{\sigma}$. However, Field's theory implies that *given a system of ordinal notations*, Ned believes that for some ordinal σ, Helen's thesis is indeterminate$_{\sigma}$. One problem with this claim is that it implies that the content of Ned's belief is relative to a system of ordinal notations. Another problem is that if Helen's thesis has a "high enough" level of indeterminacy (i.e., the ordinal it is assigned is not a member of the proper initial segment of the recursive ordinals on which the system of ordinal notations in question is defined), then the sentence Ned asserts is true, but on the AF approach, it is false. Of course, a defender of the AF approach could define a new system of ordinal notations and introduce a new hierarchy of indeterminacy predicates for a larger initial segment of the recursive ordinals with which to interpret Ned's assertion and belief, but the same problems will occur again because there is no maximal system of ordinal notations and, hence, no "maximal" hierarchy of indeterminacy predicates. Hence, this version of the AF approach does not allow Ned to believe that Helen's thesis satisfies some indeterminacy predicate or other *unless* he believes that Helen's thesis satisfies some indeterminacy predicate or other among members of S (where S is a set of indeterminacy predicates defined in terms of a fixed system of ordinal notations). Therefore, rejecting (AF5) is unacceptable.

I have addressed my criticism of hierarchy approaches to Field's theory because it is a prominent, sophisticated, and well-developed version of the hierarchy approach. Analogous criticisms can be constructed for the ambiguity versions of other hierarchy approaches.

A.7 Objections and Replies

Objection 1: One can accept a fragmentary theory of truth and accept that Ned's assertions are warranted so long as one assumes that Ned can apply the expression in question to the target while intending that the expression has the meaning of whichever restricted expression is appropriate for the target. Take the case of the ambiguity Tarskian (AT) approach as an example. Ned can assert 'Helen's thesis is true' and intend that 'true' has the meaning of the Tarskian truth predicate that is appropriate for Helen's thesis. Thus, even if he does not have a belief about the level of Helen's thesis, he can still assert that Helen's thesis is true_n and believe that Helen's thesis is true_n (for some particular n). Hence, his assertion can count as warranted despite the fact that he has no belief about the level of Helen's thesis. Similar reasoning holds for the other examples.

Reply 1: I have two replies to this objection. First, even if all the claims in the objection are true, it does not undermine the warranted assertibility argument or save the fragmentary theories of truth in question. In my example, Ned does not intend that 'true' has the meaning of whichever Tarskian truth predicate is appropriate for Helen's thesis. No such intention is required for Ned's assertion to be warranted. Thus, this version of the AT approach still implies that Ned's assertion is unwarranted. Second, the following is a brief example that illustrates a problem with this objection. Assume that Adil and Uter are having a conversation in English. Adil asserts 'at two o'clock, either I will be at the edge of the Mississippi river doing some fishing or I will be at First National depositing my paycheck'. Assume that Uter knows that Adil is a man of his word, but Uter has no belief about which action Adil will choose. After his assertion, Adil leaves, and at two o'clock, Uter encounters Lewis. Lewis utters 'where is Adil?'. In response, Uter asserts 'He is at the bank'. Assume that Uter intends the expression 'bank' in his sentence to have whichever meaning is appropriate for Adil's location. Lewis, of course, asks 'do you mean *river bank* or *financial bank*?'. In response, Uter asserts 'I do not know; I mean whichever one is appropriate'. Befuddled by Uter's shenanigans, Lewis resigns from the conversation.

Does it seem plausible to say that Uter successfully gives 'bank' in his sentence a definite meaning? Of course not. Is Uter's utterance a warranted assertion? No; in fact, I would say that it is not even an assertion.[41] We cannot expect to disambiguate our words by intending that they have whichever meaning will make the sentences we utter true. The same mistake is present in the objection. Ned cannot successfully attach a meaning to 'true' in 'Helen's thesis is true' by intending that it has whichever meaning is appropriate for Helen's thesis. In order for Ned to attach the appropriate meaning to 'true', he must have a belief about either the level of Helen's

[41] One reason for thinking that his utterance is not an assertion is that 'bank' fails to have a determinate meaning in the sentence he utters. Even if his utterance counts as an assertion, it is not warranted because it is not the case that either Uter believes that Adil is at the riverbank or Uter believes that Adil is at the financial bank.

thesis or which Tarskian truth predicate Helen's thesis satisfies. Similar reasoning holds for Ned in the other examples.[42]

Objection 2: An advocate of one of the fragmentary theories of truth criticized above can claim that, in each example, Ned has access to a definite description that picks out the relevant fact about the target of his attribution. For example, in the case of the ambiguity language-specific (ALS) approach, Ned believes that Helen's thesis is true in whichever language Helen was speaking when she asserted it. Although Ned does not know the name of that language, he possesses a definite description that picks it out. A proponent of the ALS approach can claim that Ned's sentence means that Helen's thesis is true-in-the-language-Helen-was-speaking. Thus, Ned asserts that Helen's thesis is true-in-the-language-Helen-was-speaking, and Ned believes that Helen's thesis is true-in-the-language-Helen-was-speaking; thus, his assertion is warranted. Similar reasoning holds for the other examples.

Reply 2: According to the ALS approach, 'Helen's thesis is true' is synonymous with 'Helen's thesis is true-in-L', where 'L' is a constant. If definite descriptions receive a Russellian explanation, then the sentence 'Helen's thesis is true-in-the-language-Helen-was-speaking' has the logical form: (∃x)(x is the language Helen was speaking, and Helen's thesis is true-in-x). On this account, the predicate 'true-in-the-language-Helen-was-speaking' has a hidden variable. However, LS approaches cannot appeal to a predicate 'true-in-L' where 'L' functions as a variable (see the discussion in section four). Therefore, this interpretation of Ned's sentence is not available to the ALS approach. If definite descriptions receive a more "referential" explanation, then the predicate 'true-in-the-language-Helen-was-speaking' behaves more like

[42] An example similar to the one with Adil, Uter, and Lewis figures prominently the debate between Goldberg and Brueckner. They disagree about its relevance for semantic externalism, but they agree with my conclusion; see Goldberg (1997, 1999, 2000) and Brueckner (1999, 2000).

‘true-in-L’ where ‘L’ is a name. In the reply to the next objection, I address this version of the ALS approach.

Objection 3: One can accept a fragmentary theory of truth and accept that the assertions made by Ned are warranted so long as one assumes that Ned has access to a name for the portion of discourse that determines the relevant facts about the target of the attribution in question. For example, Ned can assert ‘Helen’s thesis is true’ and intend that ‘true’ is synonymous with ‘true-in-Helen-language’ where ‘Helen-language’ is a name for the language Helen was speaking when she asserted Helen’s thesis. Ned asserts that Helen’s thesis is true-in-Helen-language, and Ned believes that Helen’s thesis is true-in-Helen-language; hence, his assertion is warranted. If ‘Helen-language’ is unavailable to Ned, he can simply create it as a name whose referent is determined by the definite description ‘the language Helen was speaking when she asserted Helen’s thesis’. Similar remarks hold for the other cases.

Reply 3: I have two replies to this objection. First, even if all the claims in the objection are true, it does not undermine the warranted assertibility argument or save the fragmentary theories of truth in question. In my example, Ned does not create a name ‘Helen-language’, nor does he intend his use of ‘true’ to be synonymous with a particular LS truth predicate. Nevertheless, Ned’s assertion is warranted, and the ALS approach implies that it is unwarranted. Second, I doubt that Ned succeeds in determining the appropriate meaning for ‘true’ in the scenario described in the objection. Consider the example with Adil, Uter, and Lewis again, but assume that they have in their language two terms ‘bank-river’ and ‘bank-downtown’ that are not ambiguous (assume for simplicity that there is only one financial bank, and it is downtown). At two o’clock, Uter creates the name ‘Adil-location’ by stipulating that it refers to the location of Adil. Uter then asserts ‘Adil is at the bank’ and intends ‘bank’ to be synonymous with ‘bank-

Adil-location'. Is his assertion warranted? No. He still has failed to give 'bank' a determinate meaning. The lesson carries over to the situation in the objection.[43]

Objection 4: The assumption that a user of an ambiguous word has to intend it to have a particular meaning is false. Speakers can use the term without implicitly attaching a meaning to it. The context in which it is used determines the meaning. This goes on "behind the speaker's back," so to speak. One can hold an analogous view for the content of the speaker's propositional attitudes. In the example for the ambiguity Tarskian approach, Ned asserts 'Helen's thesis is true' and believes that Helen's thesis is true. Ned's term 'true' takes on the appropriate meaning (e.g., if Helen's thesis is level 3, then 'true' in the sentence Ned asserts is synonymous with 'true$_3$'), and Ned's belief has the appropriate content as well (e.g., Ned believes that Helen's thesis is true$_3$). The disambiguation occurs without Ned having to do anything other than asserting 'Helen's thesis is true'. Ned's assertion is warranted even though he does not have a belief about the particular level of Helen's thesis. Similar reasoning holds for the other examples.[44]

Reply 4: One might reply to this objection by claiming that if this account of disambiguation is correct, then Ned does not know what the sentence he asserted means and does not know the content of his associated belief. Analogous criticisms are commonly made of semantic externalism (i.e., the doctrine that the contents of mental states and expressions are determined, in part, by their physical or social environment). Defenders of semantic externalism have replied to these criticisms by proposing accounts of how we know the contents of our

[43] I am indebted to conversations with Graham Hubbs and Brad Cokelet on this issue. See Kripke (1980) and Soames (2003: ch. 16) for more on these aspects of name creation and reference fixing.
[44] John Morrison suggested this objection in conversation.

mental states and expressions that are compatible with semantic externalism.[45] It seems to me that the objector could make a similar move in response to my suggested reply.

Instead of pushing that sort of reply to the objection, I want to point out what is a more pedestrian problem with it. The real issue is that the account of disambiguation endorsed in the objection is implausible. Consider once more the example with Adil, Uter, and Lewis. In this version, instead of intending anything, Uter simply asserts 'Adil is at the bank' while allowing the "context" to disambiguate 'bank' for him. This story is even less plausible than the other versions of this example. Clearly, Uter does not succeed in giving a meaning to 'bank', it is not the case that he believes that Adil is at the riverbank, and it is not the case that he believes that Adil is at the financial bank. If a person asserts a sentence that contains an ambiguous expression, he is asked the meaning of that expression on that occasion of use, and he responds to the query by saying, "I don't know," then this is a guilty admission, and a clear case where the assertion is unwarranted. Thus, the account of disambiguation in question (according to which 'true' in a speaker's assertion is disambiguated "behind his back") has the opposite problem—it treats what are clearly unwarranted assertions as warranted.[46]

There are cases that bear a certain similarity to the ones where someone asserts a sentence without knowing the meaning of an ambiguous word in the sentence. Consider Cletus, who asserts 'the bird is red'. Assume that Cletus's wife, Brandine, challenges him by asking, "By 'red' do you mean *brick*, *crimson*, *maroon*, etc.?" If Cletus says, "I don't know," in response to Brandine's query, but refuses to retract his assertion, then the most natural way of

[45] The literature on semantic externalism is vast. See Putnam (1975), Burge (1979b), and Davidson (1988) for arguments in favor of semantic externalism. For criticism see Boghossian (1989) and Segal (2000). For attempts to reconcile semantic externalism with knowledge of the contents of one's expressions and mental states, see Davidson (1987), Burge (1988), Kobes (1996), Gibbons (1996), and Heal (2001).

[46] One should not assume that this reply casts doubt on semantic externalism. There is an important difference between using 'water' while being secretly switched from Earth to Twin Earth and using a word one knows to be ambiguous without intending any particular meaning. It is akin to the difference between confusion and ambiguity; see Camp (2002: ch. 5).

understanding Cletus's original assertion is *not* that 'red' is ambiguous and that the "context" disambiguates it "behind Cletus's back." The obvious interpretation is that 'red' is being used as a *generic* predicate. Cletus's claim means something like: there is some kind of red such that the bird is that kind of red. There is a moral for proponents of fragmentary theories of truth: if a language contains a set of expressions for concepts that fall under an intuitive kind, then that language should either have a generic term for them or allow for introduction of a generic term for them.[47]

Objection 5: There are theories of warranted assertibility on which someone who asserts that p need not have the *de dicto* belief that p for her assertion to count as warranted. One might take *de re* belief to be enough, or one might claim that truth is the only norm of assertion.[48] On such theories of warranted assertibility, the fragmentary theories of truth considered above do not imply that Ned's assertions are unwarranted.

Reply 5: Even if it turns out that the best theory of warranted assertibility does not require the asserter to believe (*de dicto*) what he or she asserted, it will still have to respect Keith DeRose's distinction between primary and secondary propriety.[49] If S believes that he is acting in accordance with the norms that govern warranted assertibility, then, although S's assertion might be unwarranted, S is still acting properly in the sense that he should not be sanctioned for his action by the members of his community. Call such an assertion *responsible*. Even if the

[47] One might be tempted to offer a similar objection based on disambiguation via the division of linguistic labor (see Putnam 1975). The objection would be something like: because the asserter lives in a community in which there are people who know the relevant facts about the target of Ned's attribution, he can assert a sentence with the ambiguous word and allow the opinions of the experts to disambiguate it for him. My reply is similar in spirit. First, there is no guarantee that a member of Ned's community will have the relevant knowledge; if no such person is available, then his assertion is still unwarranted. Second, if Ned does not know that the word is ambiguous, then he is confused (he thinks that two or more distinct entities are identical); confusion cannot be explained in terms of ambiguity; see Camp (2002: ch. 5). If Ned does know that the word is ambiguous, and he does not intend that it has one meaning rather than another, then either he uses it as a generic word (which is not ambiguous) or he does not know what the word means in his utterance. It makes no difference whether he lives in a community of people who know what he should mean in that circumstance in order to make the sentence he asserts true.

[48] See Williamson (2000a: ch. 11) for discussion.

[49] DeRose (2002: 180).

primary propriety for assertibility is bound up with *de re* belief or the truth of the sentence asserted, the secondary propriety for assertibility depends on the asserter's *de dicto* belief that p when asserting that p. If S asserts a proposition, S's assertion is responsible (S made a sufficient effort to follow the relevant norms of assertion), and the relevant norms involve either the truth of the sentence asserted or *de re* belief in the proposition it expresses, then S has the *de dicto* belief in that proposition. The warranted assertibility argument goes through with 'responsible assertion' in place of 'warranted assertion'. The resulting argument shows that the fragmentary theories of truth I discuss do not respect our intuitions about responsible assertibility; a theory that implies that some responsible assertions are irresponsible is just as bad as one that implies that some warranted assertions are unwarranted.

Objection 6: There are other ways for a proponent of a fragmentary theory of truth to explain the relation between the restricted expressions posited by the theory and the expressions of natural language. For example, one can identify the natural language expression with one of the restricted expressions and introduce the other restricted expressions into the language. Another option is to assume that the expression of natural language is context-dependent.

Reply 6: I stated in section one that I would not consider context-dependence versions of fragmentary theories of truth. As I have presented it here, the warranted assertibility argument might seem to lend support to context-dependence versions because it casts doubt on the alternatives; I want to caution against this assessment. I argue elsewhere that context-dependence approaches to the liar paradox are untenable, but I do not have the space to recreate that argument here.[50] As for the other suggestion in the objection, no account of that sort will be a plausible descriptive theory of truth. Consider the example of the ambiguity Tarskian approach. If one were to stipulate that 'true' of English is synonymous with 'true$_0$', then that

[50] See Appendix B.

theory would imply that ‘p is true’ is false whenever p contains a truth predicate. That account is obviously false. Similar results hold for the other fragmentary theories of truth when interpreted in this way. Of course, there are other ways to interpret fragmentary theories of truth, but the warranted assertibility argument shows that if they are to be acceptable, they must be consistent with our intuitions about which assertions of truth attributions are warranted.

Objection 7: Perhaps most language users would say that the assertions performed by Ned are warranted. However, they should not count as warranted. Those sorts of assertions happen infrequently, and changing our linguistic practice so that they do not count as warranted would have a minor impact. That is a small price to pay for being able to use the fragmentary theories of truth considered here.

Reply 7: The objection suggests that we treat fragmentary theories of truth as revisionary theories of truth—ones that specify how we *should* think and talk about truth instead of ones that purport to describe our actual linguistic practice. There are two problems with this suggestion. First, if we want to understand our current linguistic practice, then the fragmentary theories of truth I consider are not going to be of help. That is the conclusion of the warranted assertibility argument. Reinterpreting fragmentary theories of truth as revisionary theories does nothing to change that fact.

Second, the ability to blindly assert truth attributions that count as warranted is essential to the functioning of a truth predicate (a *blind* assertion of a truth attribution occurs when the asserter is not in a position to assert the targets of the attribution). Deflationists and non-deflationists alike can agree on this issue. If truth is a substantive property (as a non-deflationist holds) and a person has good reason to believe that some particular sentences are true, then she should be able to attribute truth to them in a warranted assertion without having to believe other things about them (e.g., which levels they have, which languages they belong to, or whether they are translatable into the language she is using). Likewise, deflationists have good reason to permit blind assertions of truth attributions; otherwise they lose the deflationist account of the value of truth predicates, which is that truth predicates allow us to make generalizations we could not otherwise make.[51] Without the ability to assert truth attributions that are both blind and warranted, that would be impossible. Thus, a concept of truth offered by a revisionary theory of truth that abandons this aspect of our use of 'true' would be of little use to us.[52]

A.8 Conclusion

I have argued that certain fragmentary theories of truth are unacceptable. These include the ambiguity version of Tarski's theory of truth, the ambiguity language-specific approach, the translational language-specific approach, and the ambiguity version of Field's theory of truth and indeterminacy (I also suggested that the criticism of Field's theory applies to the ambiguity versions of other hierarchy approaches). In each case, I used the warranted assertibility argument to show that the theory in question is inconsistent with our intuitions about which assertions are warranted.

[51] See Field (1994a), Horwich (1998), and Halbach (1999).

[52] Despite Field's claims to the contrary (See Field 2003c: 300), it is not plausible to interpret his theory as a descriptive theory of truth and a revisionary theory of indeterminacy. For Field, if p is a sentence of a language L, and p is a member of neither the extension nor the anti-extension of 'true-in-L', then neither 'p is not true-in-L' nor 'p is not false-in-L' are in the extension of 'true-in-L'. Instead, Field delegates this traditional role of 'true' to his indeterminacy predicates, each of which is built up from conjunction, (choice) negation, truth predicates, and determinacy operators. When one applies Field's theory to English, one can either (i) stipulate that 'true' can be correctly applied to gappy sentences, but 'true' is ambiguous and can have the meaning of any of the determinate truth predicates, or (ii) stipulate that 'true' cannot be correctly applied to gappy sentences. In the first case, a version of the warranted assertibility argument shows that Field's theory implies that some warranted assertions are unwarranted; in the second, the theory is a revisionary theory of truth. Again, we have the option of a false descriptive theory or a revisionary theory.

Deflationism and approaches to the liar paradox provide the motivation for fragmentary theories of truth. I have not argued that deflationism or these approaches to the liar paradox are false; nor have I demonstrated that they have nothing to do with natural language truth predicates. Rather, I have shown that if they are relevant to natural language truth predicates, then the proponents of these theories must provide some account of the relation between the restricted truth predicates to which they appeal and natural language truth predicates that does not run afoul of the warranted assertibility argument.

The motivations for fragmentary theories of truth are powerful, and I have not shown that all fragmentary theories of truth are unacceptable. Indeed, I advocate a fragmentary theory of truth. However, it is neither a language-specific approach nor a hierarchy approach. Rather, it appeals to a group of six restricted truth predicates. Instead of treating natural language truth predicates as ambiguous or context-dependent, I claim that they are confused. The approach I offer permits a generic truth predicate, and it respects our intuitions about which assertions of truth attributions are warranted. It also solves the liar paradox without relying on a hierarchy of semantic predicates, a substantive distinction between object language and metalanguage, or a restriction to expressively weak languages.

APPENDIX B

Risky Business: Truth and Paradoxicality

B.1 Introduction

Some sentences of natural languages that contain truth predicates are paradoxical. By 'paradoxical', I mean that one can prove that these sentences are both true and not true from some intuitively plausible assumptions via intuitively plausible inference rules. I do *not* mean that these sentences *are* both true and not true. The most well-known example of a paradoxical sentence is the liar sentence:

(Λ) λ is false.

A sentence token of type Λ whose name is 'λ' attributes falsity to itself; such a sentence is paradoxical.[1] It is common knowledge among those who work on the liar paradox that one can construct paradoxical sentences with the use of empirical predicates. These sentences are paradoxical because of some empirical facts; if the facts had been different, they would not have been paradoxical. We can say that such sentences are *empirically paradoxical*. The existence of

[1] The argument is based on the *truth rules* (i.e., ⟨⟨p⟩ is true⟩ follows from ⟨p⟩ and vice versa, and two names that refer to ⟨p⟩ are intersubstitutable in ⟨⟨p⟩ is true⟩ without changing the truth-value of the sentence). On the one hand, if λ is true, then 'λ is false' is true. If 'λ is false' is true, then λ is false. Thus, if λ is true, then λ is false. On the other hand, if λ is false, then 'λ is false' is true. If 'λ is false' is true, then λ is true. Thus, if λ is false, then λ is true. Therefore, λ is true if and only if λ is false. It follows that λ is both true and false.

these sentences suggests that a sentence's syntactic and semantic features do not determine whether it is paradoxical. In section two, I discuss this intuition, and in section three, I provide an argument for it.

Although many contemporary philosophers who work on truth pay lip service to empirical paradoxicality, few realize that it has sweeping consequences for a wide range of views related to truth. In particular, the thesis that for some sentences, their syntactic and semantic features do not determine whether they are paradoxical is *incompatible* with (i) the claim that utterances of paradoxical sentences do not count as assertions, (ii) theories that make propositions the primary bearers of truth and falsity, (iii) minimalist accounts of truth-aptness, which hold that the syntactic features of a sentence determine whether it has a truth-value, (iv) the version of deflationism that defines truth predicates in terms of sets of T-sentences (e.g., ''spandex jumpsuits are hot' is true if and only if spandex jumpsuits are hot'), and (v) contextualist theories of truth, which seek to avoid the liar paradox by stipulating that the extensions and anti-extensions of truth predicates vary from context to context. I develop these criticisms in sections four through nine.

B.2 RISKINESS: INTUITION

Philosophers have known of empirical versions of the liar paradox since it first became an object of study over two millennia ago. For example, the predicate 'is a complete sentence in section two of Scharp's "Risky Business," whose first letter is an 'E'', can be used to construct a version of the liar paradox; consider the following sentence.

> Every complete sentence in section two of Scharp's "Risky Business," whose first letter is an 'E' is false.

The fact that the previous sentence is the only complete sentence in section two of this appendix to begin with an ‘E’ is an empirical fact about that sentence. If I had chosen to place it in a different section, it would not have satisfied that empirical predicate and, thus, it would not have predicated falsity of itself.[2] Nevertheless, it seems obvious that this change would not have altered the sentence’s syntactic features or its meaning.

Although empirical versions of the liar paradox are as old as the paradox itself, they have not received much attention in contemporary discussions.[3] Certainly the most influential examination of empirically paradoxical sentences is found in Kripke’s paper on truth.[4] Kripke draws a striking conclusion from the fact that there are empirical versions of the liar paradox:

> The versions of the Liar paradox which used empirical predicates already point up one major aspect of the problem: *many, probably most, of our ordinary assertions about truth and falsity are liable, if the empirical facts are extremely unfavorable, to exhibit paradoxical features*, (Kripke 1975: 691; italics in original).

He provides his own example of this phenomenon in which Nixon and Jones assert sentences that, due to their satisfaction of empirical predicates, are paradoxical. Kripke claims that his example “points up an important lesson: it would be fruitless to look for an *intrinsic* criterion that will enable us to sieve out—as meaningless, or ill-formed—those sentences which lead to paradox,” (Kripke 1975: 692). He continues, “The moral: an adequate theory must allow our statements involving the notion of truth to be *risky*: they risk being paradoxical if the empirical facts are extremely (and unexpectedly) unfavorable. There can be no syntactic or semantic ‘sieve’ that will winnow out the ‘bad’ cases while preserving the ‘good’ ones,” (Kripke 1975:

[2] I assume that sentence tokens are truth bearers for reasons I discuss in section five.

[3] See Church (1946), Cohen (1957, 1960), Prior (1958, 1961), van Fraassen (1968), Burge (1979), Gupta (1982), Yablo (1982), Martinich (1983), Parsons (1984), Barwise and Etchemendy (1987), Kremer (1988), Stebbins (1992), Gaifman (1992), Simmons (1993: ch. 8), Goldstein (2001), and Visser (2004) for remarks on empirical paradoxicality.

[4] Kripke (1975).

692). In these passages, Kripke claims that: (i) the existence of empirically paradoxical sentences shows that approaches to the liar paradox on which paradoxical sentences are either ungrammatical or meaningless are unacceptable, (ii) a theory of truth should imply (or at least permit) that truth attributions are risky, and (iii) it is not the case that for every sentence token σ, the syntactic and semantic features of σ determine whether σ is paradoxical. (i) follows from (iii), and the most natural reading of (ii) is that a theory of truth should imply (iii). Thus, (iii) seems to be the most important of these claims. I refer to it as the *riskiness thesis*. However, given the significance of this claim (which I draw out in sections five through nine), I would prefer a more convincing justification of it.

What is needed is an example with two situations, one in which a person asserts a token of a sentence type that is not paradoxical and the other in which a person asserts a token of the same sentence type that is paradoxical. In both situations, the people are molecule for molecule identical, they have the same mental states, the sentences have the same meanings, their subsentential parts have the same meanings, and the singular terms that occur in the sentences refer to the same things. I provide such an example in the next section. Before doing so, I want to emphasize that I am *not* arguing that the semantic and syntactic features of a sentence token never determine whether it is paradoxical. Indeed, the syntactic and semantic features of some sentence tokens do determine whether they are paradoxical.[5] I am arguing that for some sentence tokens, their syntactic and semantic features do not determine whether they are paradoxical.

[5] Tarski (1944) argued for this claim. See also Quine (1961) in which the example ‘‘yields a falsehood when appended to its own quotation’ yields a falsehood when appended to its own quotation’ features prominently.

B.3 Riskiness: Example

Let w_1 and w_2 be two possible worlds that are very much like our own (e.g., they have the same natural laws[6], they have very similar histories, humans exist in both, some humans speak English in both, etc.). I am appealing to possible worlds at this point because they provide a relatively easy way to make my point. The conclusions I draw from this example hold across a range of views on the role of 'possible worlds' talk.

In w_1 there is a person, Stu, who is in a room that is empty except for a single blackboard and another person, Gil.[7] Stu is a competent English speaker and all the sentences in the example are sentences of English (my idiolect at noon GMT on January 1, 2004, to be precise). Stu asserts a token of the following sentence type:

(A) The sentence written on the blackboard is true.

Let 'α' be the name of the sentence token Stu utters. Thus, α is a sentence token of sentence type A. Assume that Stu indicates by gesture that the sole blackboard in the room is the one to which he is referring. Assume as well that there is a single sentence written on that blackboard. The sentence token written on the blackboard is a token of the following sentence type:

(B) The sentence written on the blackboard in the room next door is true.

Let 'β' be the name of the sentence token that is written on the blackboard in the room in which Stu and Gil are present. For simplicity, I call the room in which Stu and Gil are present, *room 1*. Assume also that there is only one room next door to room 1. I call it *room 2*. Assume that

[6] Advocates of Humean supervenience (i.e., the natural laws of a world supervene on its non-nomic facts) should feel free to assume that w_1 and w_2 have very similar natural laws, but because their non-nomic facts differ slightly, their natural laws will as well. This complication makes no difference for my argument.

[7] At this point, Gil is present merely to insure that Stu has an audience so that his utterances count as assertions; in section eight, Gil plays a more substantial role.

room 2 is empty except for a single blackboard and that on that blackboard is a single sentence. The lone sentence on the lone blackboard in room 2 is a token of the following type:

(Γ_1) β is false.

Let 'γ_1' be the name of the sentence token written on the blackboard in room 2 in world w_1. Thus, γ_1 is a sentence token of the sentence type Γ_1.

In w_1, α, β, and γ_1 are paradoxical. An argument similar to the one used to show that λ is paradoxical is sufficient to demonstrate this fact.[8]

For my purposes, I do not need to provide a rigorous definition of 'paradoxical'. Or, better, any of the rigorous definitions of paradoxicality in the literature would serve my purposes.[9] Roughly, a sentence token is paradoxical if and only if from the assumptions that $\langle\langle p\rangle$ is true$\rangle$ and $\langle p\rangle$ are intersubstitutable in extensional contexts, that '___ is true' is an extensional context, and that the sentence token is either true or false, we can derive that the sentence token is both true and false using the inference rules of classical logic.

World w_2 is exactly like w_1 in every detail except that in w_2, a different sentence token is inscribed on the blackboard in room 2.[10] In world w_2 on that blackboard in room 2 is a token of the following sentence type:

(Γ_2) Monkeys like bananas.

[8] See footnote 1.

[9] See Chihara (1973) for an account of paradoxicality in terms of the vicious circle principle. For an account of paradoxicality in terms of diagonalization, see Thomson (1962), Richards (1967), Goddard and Johnston (1983), Goddard (1984), and Simmons (1990, 1993); see Mackie (1973), Chihara (1973), and Martin (1976) for criticism. See Tennant (1982, 1995) for a proof-theoretic characterization of paradoxicality. See Priest (1994) for an account in terms of the qualified Russell schema; see Grattan-Guinness (1998), Priest (1998), N. Smith (2000) for discussion. Most of the approaches to the liar paradox include accounts of paradoxicality as well. See Kripke (1975) for a fixed-point characterization, Yablo (1985) for a stage-theoretic characterization, Gupta and Belnap (1993) for a revision-theoretic characterization, and McDonald (2000) for a variational characterization. See also McGee (1991), Simmons (1993), and Field (2003b).

[10] Of course, that cannot be the only difference, but I am ignoring backtracking issues; see Lewis (1979b: 32-35).

Let 'γ_2' be the name of the sentence token on the blackboard in room 2 in world w_2. In world w_2, α, β and γ_2 are not paradoxical. If we assume that monkeys do indeed like bananas and that γ_2 is truth-apt, then α, β, and γ_2 are all true in w_2.

The example shows that paradoxicality does not supervene on the syntactic and semantic features of a sentence token. In w_1 and w_2, Stu and Gil are molecule for molecule identical, they have the same histories, and they have the same mental states. Thus, the syntactic and semantic features of α that depend on who produced α are the same in w_1 and w_2. Indeed, in w_1 and w_2, α has all the same syntactic and semantic features. In each world, it belongs to the same language, it is a closed, well-formed, declarative sentence, it has the same grammatical structure, and it has the same subsentential parts. In w_1 and w_2, α has the same sentential meaning, its subsentential parts have the same subsentential meanings, its predicates have the same extensions, and its singular terms have the same referents. Yet in w_1, α is paradoxical, and in w_2, α is not paradoxical.

Before leaving this example, I want to address several potential objections. First, there are views according to which α has different sentential meanings in w_1 and w_2. For example, according to one version of the Tarskian approach to the liar paradox, the truth predicate in α in w_1 is different from the truth predicate in α in w_2. The Tarskian approach to the liar paradox appeals to a hierarchy of truth predicates, 'true_0', 'true_1', etc. The "lowest" truth predicate, 'true_0', applies only to sentences that do not contain truth predicates. 'True_1' applies only to sentences that contain 'true_0' but no other truth predicates, 'true_2' applies only to sentences that contain 'true_1' or 'true_0', but no other truth predicates, etc. There are several ways of using the Tarskian hierarchy to explain a natural language truth predicate. One can stipulate that a natural language truth predicate is ambiguous and that it can take on the meaning of any one of the truth predicates in the hierarchy.[11] An advocate of this view might say that the meaning of the natural language truth predicate is determined not by the intentions of the speaker, but by the target of the attribution. If no predicate of the Tarskian hierarchy is appropriate for the target of the attribution, then 'true' in the sentence in question is meaningless. On this theory of truth, 'true' in α in w_1 is synonymous with 'true$_1$', while 'true' in α in w_2 is meaningless. Thus, according to this view, α has different semantic features in the two worlds.

Contextual theories of truth have this consequence as well.[12] These views hold that natural language truth predicates are context dependent. The contextual approaches stipulate that the extension of a natural language truth predicate changes from context to context in such a way that sentences that would be paradoxical are excluded from the extension. Theories of context dependent expressions commonly distinguish between an expression's meaning, which is constant, and its content, which varies. On the contextual theory, although the occurrence of 'true' has the same *meaning* in w_1 and w_2, it has different *content* and a different extension in the two worlds. Depending on the particular theory, sentences that seem paradoxical are either false in the context or gappy in the context.

In my view, there are serious problems with both the claim that natural language truth predicates are ambiguous and the claim that they are context dependent. I discuss the former

[11] Toms (1956), Wormell (1958), Huggett (1958), Whiteley (1958), Herzberger (1966), and Williamson (2000a) for advocates of this view. Kripke attributes it to Parsons (1974) as well in Kripke (1975: 695 n. 10). Kripke criticizes this approach to natural language truth predicates in Kripke (1975) and I develop that criticism in Appendix A.

[12] Parsons (1974, 1983, 1984), Thomason (1976), Burge (1979a, 1982a, 1982b), Gaifman (1982, 1992, 2000), Hodges (1986), Barwise and Etchemendy (1987), Koons (1992, 2000b), Simmons (1993, 2000, 2003), Cantini (1995), and Glanzberg (2001, 2004). For criticism of Burge, see Gupta (1982) and Simmons (1993); for criticism of Gaifman see Simmons (1993) and Yi (1999); for criticism of Barwise and Etchemendy, see Gupta (1989), Grim (1991), McGee (1991), Gaifman (1992), and Priest (1993); for criticism of Simmons, see Antonelli (1996), Hardy (1997), and Beall (2003). For discussion of revenge liars for contextual theories see Hazen (1987), Hinckfuss (1991), Juhl (1997), Clark (1999), Weir (2000, 2001), and Weir (2002).

elsewhere[13] and the latter below. For now, I assume that truth predicates are univocal and invariant.

Another potential objection is that α in w_1 and α in w_2 are tokens of different types.[14] I do not want to enter in the debate about what makes a given object a token of a certain type. I assume that in ordinary situations a competent human who comprehends a language L can determine whether a given object is a sentence token of L and can determine whether any two sentence tokens of L are tokens of the same sentence type. That rules out views that treat α in w_1 and α in w_2 as tokens of different types.

B.4 Paradoxicality

Although a survey of all the approaches to the liar paradox that have been suggested in the last century is beyond the scope of this paper, I can briefly describe the major views on the status of paradoxical sentences. One view is that paradoxical sentences are not syntactically well-formed.[15] No one has defended this account in print in decades. The most likely reason for this dearth of proponents is that paradoxical sentences seem to be well-formed and the commonly accepted rules of grammar imply that they are well-formed. If that is not enough, the example in section two shows that if paradoxical sentences are not well-formed, then whether a sentence is well-formed can depend on unrelated empirical facts (in the example, whether α is well-formed would depend on which sentence token is written on the blackboard in room 2). I assume that if a competent person comprehends a language, then that person can determine whether a given

[13] See Appendix A.

[14] On the distinction between types and tokens, see Kaplan (1973), Szabó (1999), and Truncellito (1999). See Kaplan (1990) for criticism.

[15] Jørgensen (1953). See Kattsoff (1955) for criticism.

syntactic string is a well-formed sentence of that language solely on the basis of *local inspection* (i.e., by looking at it, listening to it, touching it, tasting it, etc.) and *linguistic investigation* (i.e., consulting a dictionary, a thesaurus, a grammar book, etc.). Thus, whether a sentence is well-formed cannot depend on arbitrary empirical facts.

A similar view is that paradoxical sentences are meaningless.[16] This approach was popular in the first half of the twentieth century; it has since fallen out of favor and has been recently endorsed only by a handful of deflationists who claim that all ungrounded sentences are meaningless because of the way truth predicates function.[17] There are several ways to define 'grounded', but the most intuitive is the following. One begins with sentences that do not contain 'true'; these are level 0 sentences. One then constructs all the possible sentences that attribute truth to level 0 sentences; these are level 1 sentences. One then constructs all the possible sentences that attribute truth to level 1 sentences; these are level 2. And so on. If a sentence is assigned a level in this way, then it is grounded. Otherwise, it is ungrounded. Note that all paradoxical sentences are ungrounded, but some ungrounded sentences are not paradoxical (e.g., 'no sentence is both true and false').[18]

One problem with the claim that all paradoxical sentences are meaningless is that paradoxical sentences seem to be meaningful (i.e., one can have the impression that one understands them) and they seem to have the properties that meaningful sentences have (e.g., they seem to participate in inferences, it seems that one can use them to express beliefs, etc.). Moreover, the example in section two shows that if paradoxical sentences are meaningless, then whether a sentence is meaningful can depend on unrelated empirical facts (in the example in section two, whether α is meaningful would depend on which sentence token is written on the

[16] Skinner (1959), Ross (1969); for criticism see Popper (1954), Rozeboom (1957), and Skyrms (1970).
[17] See Grover (1976, 1977), Brandom (1994: ch. 5); see also Beall (2001), and Armour-Garb (2001) for discussion.
[18] See Herzberger (1970), Kripke (1975), Yablo (1982), and McCarthy (1988) for more on grounding.

blackboard in room 2). I assume that if a person comprehends a language, then that person can determine whether a given well-formed sentence of that language is meaningful on the basis of local inspection and linguistic investigation alone. Thus, whether a sentence is meaningful does not depend on arbitrary empirical facts.[19]

Some hold that paradoxical sentences are false. For example, one version of the Tarskian approach holds that paradoxical sentence tokens attribute truth or falsity using a truth predicate of level n to a sentence of level n. Any sentence that has that feature is false.[20] Another view is that every proposition entails the conjunction of it and the proposition that it is true. According to this view, the propositions expressed by paradoxical sentences are simply contradictions.[21] Thus, paradoxical sentences are false.

The most popular and influential view is that paradoxical sentences are truth-value gaps.[22] That is, paradoxical sentences are in neither the extension nor the anti-extension of 'true'. There are many different theories that have this consequence and equally many philosophical justifications for treating paradoxical sentences as truth-value gaps. There are also several different views on what truth-value gaps are. Some take them to be simply the lack of a truth-value. Others take them to be a different kind of truth-value. I take no position on these issues.[23, 24]

[19] Therefore, the deflationist theories of truth that imply that ungrounded sentences are meaningless are unacceptable.

[20] Tarski (1933); see Church (1976), Halbach (1995), Soames (1999), and Glanzberg (2005) for discussion. For criticism see Kripke (1975), Simmons (1993), and Appendix A.

[21] Ushenko (1937), Michael (1975), and Mills (1998). I am not aware of any commentary on these approaches.

[22] See Ryle (1951), Fitch (1964), Martin (1967), van Fraassen (1968), Skyrms (1970), Kripke (1975), Feferman (1982), Reinhardt (1986), McGee (1991), Soames (1999), McDonald (2000), Blamey (2002), Field (2003a, 2003b), and Maudlin (2004). For criticism see Simmons (1993), Gupta and Belnap (1993), and Glanzberg (2003).

[23] See Kijania-Placek (2002) and Blamey (2002).

[24] In the following passage, Kripke argues that the riskiness thesis implies that the gap approach to the liar paradox is the only acceptable one: "The orthodox assignment of intrinsic levels guarantees freedom from "riskiness" in the sense explained in sec. I above. For (4) and (5) below, the very assignment of intrinsic levels which would eliminate their riskiness would also prevent them from "seeking their own levels" (see pp. 695-697). *If we wish to allow sentences to seek their own levels apparently we must also allow risky sentences.* Then we must regard sentences as

A radical view is that paradoxical sentences are both true and false. The approach with this consequence is called *dialetheism*.[25] Because most philosophers agree that it is constitutive of the concepts of truth and falsity that a sentence token cannot be simultaneously both true and false, most philosophers reject dialetheism. However, when rejecting dialetheism, one must be careful to avoid saying that dialetheism is false. It turns out that the most natural version of dialetheism implies that the sentences that constitute dialetheism are both true and false. Thus, informing a dialethist that his theory is false is not a way of rejecting his theory (according to the dialetheist). Not only can one be a dialetheist and accept that dialetheism is false, a dialetheist must accept that it is false. Indeed, the fact that dialetheism is false is the one thing that everyone (dialetheists and non-dialetheists alike) can agree on. I follow the vast majority of philosophers in rejecting dialetheism.[26]

One final take on the status of paradoxical sentences I mention comes from the revision theory of truth, according to which truth is a circular concept. The details of this subtle and ingenious theory are not relevant to my argument here. It is sufficient to say that, according to

attempting to express propositions, and allow truth-value gaps," (Kripke 1975: 695 n. 10). I find this argument intriguing, and I am unaware of any attempt to formulate it rigorously, but I do not attempt to do so here.

[25] See Priest (1979, 1987, 1998), and Armour-Garb and Beall (2001). For criticism see Parsons (1990), Everet (1996), Bromond (2002), Shapiro (2002), and Field (forthcoming a). Dialetheism is sometimes characterized by its adherents as the view that some contradictions are true. This characterization seems unfortunate to me. One should distinguish the doctrine that implies some instances of the negation of the principle of non-contradiction from the doctrine that some sentences are both true and false. One can accept one of these doctrines and reject the other. The distinction is similar to the distinction between the rejection of the principle of excluded middle and the rejection of bivalence. Strictly speaking, dialethists do not claim that some truth bearers are both true and false. Rather, they advocate three-valued logics whose third value is designated (and so interpreted as both truth and falsity).

[26] My views on justifying the rejection of dialetheism are similar to David Lewis's, which he summarizes in the following passage:

> The reason we should reject [dialetheism] is simple. No truth does have, and no truth could have, a true negation. Nothing is, and nothing could be, literally both true and false. This we know for certain, and apriori, and without any exception for especially perplexing subject matters. … That may seem dogmatic. And it is: I am affirming the very thesis that Routley and Priest have called into question and – contrary to the rules of debate – I decline to defend it. Further, I concede that it is indefensible against their challenge. They have called so much into question that I have no foothold on undisputed ground. So much the worse for the demand that philosophers always must be ready to defend their theses under the rules of debate, (Lewis 1982: 101).

That is, a non-dialetheist is entitled to ignore dialetheism.

the revision theory of truth, it is inappropriate to assert that paradoxical sentences are true and inappropriate to assert that they are false, but that does not imply that paradoxical sentences are truth-value gaps. According to the revision theorist, truth and falsity are not up to the task of characterizing paradoxical sentences. Instead, one can say that paradoxical sentences are not categorically true and not categorically false. Most important for my purposes is that it is not the case that the revision theory implies that paradoxical sentences are true.[27]

That covers the major views on the status of paradoxical sentences. No approach to the liar paradox implies that paradoxical sentences are true. That claim is important enough that you should read it again, this time in italics: *no approach to the liar paradox implies that paradoxical sentences are true*. As I have said, some contextualist theories imply that paradoxical sentences are true in some contexts and false or gappy in others, but that is quite different than claiming that paradoxical sentences are true. I am setting these approaches aside for the moment. Dialetheism implies that paradoxical sentences are both true and false, but, again, that is different from claiming that they are true (and not false). I am not considering dialetheism in this paper.

B.5 Assertion

In this section, I use the riskiness thesis to criticize the view that utterances of paradoxical sentences do not count as assertions. This view is espoused explicitly by some who use it as an approach to the liar paradox.[28] It also follows from several prominent views on assertion and the liar paradox. For example, Glanzberg argues that utterances of sentences that are truth-value

[27] See Gupta (1982, 1997, 2002), Herzberger (1982a, 1982b), Gupta and Belnap (1993), Yaqūb (1993), and Chapuis (1996). For criticism see Hart (1989), Simmons (1993), Koons (1994), McGee (1997), Martin (1997), and Cook (2002).

[28] See Prior (1958, 1961), Richards (1967), Martinich (1983), and Goldstein (1991, 1992, 1999, 2001) for examples of this approach.

gaps do not count as assertions.[29] When coupled with the most prominent approach to the liar paradox (the truth-value gap approach), it follows that utterances of paradoxical sentences are not assertions. In addition, some philosophers explain assertion as a propositional attitude.[30] When this view is combined with the claim that paradoxical sentences do not express propositions, it follows that utterances of paradoxical sentences are not assertions.[31] Thus, the claim that paradoxical sentences cannot be asserted follows from some pretty common views. I argue that it is incompatible with the riskiness thesis.

I have two objections. First, we do not treat utterances of paradoxical sentences as if they are not assertions. Evidence for this claim comes from our practice of warrantedly asserting assertion attributions to blindly uttered truth attributions. That is, we assert that a particular utterance of a truth attribution (i.e., a sentence of the form: ⟨p⟩ is true) is an assertion even if we do not know what the target of that truth attribution is and we know that the person who uttered it does not know what the target is either. For example, on Tuesday at noon, Bob blindly utters a truth attribution, ‘the sentence Merle uttered yesterday at noon is true’. Bob does not know which sentence Merle uttered yesterday at noon, but Bob has been informed by someone he trusts that the sentence is true. If the sentence in question is ‘the sentence Bob will utter tomorrow at noon is false’, then both sentences are paradoxical. On the view in question, if that is indeed the sentence Merle uttered on Monday at noon, then neither Merle’s utterance nor Bob’s utterance counts as an assertion. Assume that Cecil is in Bob’s presence at noon on Tuesday and hears his utterance. Cecil utters ‘Bob’s utterance is an assertion’. Cecil does not know the identity of Merle’s utterance either, but he is justified in attributing assertionhood to Bob’s utterance. This is an established practice. We simply do not treat paradoxical utterances

[29] Glanzberg (2003); see also Stalnaker (1978).
[30] Stalnaker (1970, 1974, 1978, 1998) and Soames (2002).
[31] See Kripke (1975) for the claim that paradoxical sentences do not express propositions.

as if they fail to be assertions. Thus, the claim that utterances of paradoxical sentences are not assertions is best interpreted as a revisionary suggestion, not as a descriptive claim. That is, it is a claim about how we *should* conduct our linguistic practice, not a claim about how we actually behave.

Given that the claim in question fails as a descriptive theory, we should ask: should we use 'assertion' in the way it suggests? I argue that we should not. The problem, in essence, is that if utterances of paradoxical sentences failed to be assertions, then whether an utterance of a sentence that contains 'true' counts as an assertion could depend on any arbitrary fact. This is at best a counterintuitive result, and at worst an unacceptable one. I claim that it is unacceptable because it implies that humans in everyday discourse situations cannot in general determine whether an utterance of a sentence containing a truth predicate is an assertion. The view that humans cannot in general determine the pragmatic force of utterances of sentences in which a certain linguistic expression occurs is unacceptable. I cannot adequately defend such a big claim in such a small space, but I do want to provide the sketch of an argument for it.

Here is the argument (which I call the *availability argument*):

(1A) Paradoxicality (i.e., whether a sentence token on an occasion of use is paradoxical) need not be available to the participants in a conversation.

(2A) Assertionhood (i.e., whether an utterance of a sentence token is an assertion) is available to participants in a conversation.

∴ (3A) Assertionhood does not depend on paradoxicality (i.e., whether an utterance of a sentence token is an assertion does not depend on whether that sentence token is paradoxical).

Defending the first two premises requires an account of what is available to participants in a conversation. I use Stalnaker's theory of conversational context for this purpose (I have no attachment to Stalnaker's theory other than that it should be familiar to many readers; I assume that other accounts would work just as well).

On Stalnaker's view, at each stage in a conversation, the participants have certain *presuppositions*, which are treated as propositional attitudes, not as semantic relations. Both the participant's presuppositions and the propositions presupposed are modeled on sets of possible worlds. A proposition is modeled on the set of possible worlds in which it is true. Presuppositions are modeled on a set of possible worlds in which the proposition presupposed is true. The intersection of the sets of possible worlds associated with a participant's presuppositions is called the *context set*. In nondefective conversations, the context sets of the participants are identical. Thus, in such conversations, the context of the conversation is just the context set of each participant. In defective conversations, the context is not well-defined. An *assertion* reduces the context set by eliminating the possible worlds in which the proposition asserted is false. The context is available to the participants of the conversation in the sense that they know what their presuppositions are and so they know what their context sets are. So long as the conversation is nondefective, they know what the context is.[32]

Premise (1A) is that paradoxicality need not be available to the participants of a conversation. This claim should be obvious from the example in section two. In the conversation between Stu and Gil, the context is the same in world w_1 and world w_2. However, α is paradoxical in w_1 and not paradoxical in w_2. Thus, paradoxicality need not be determined by the conversational context.

Defending premise (2A) is more complicated. To begin, I want to consider how the force of an utterance affects a conversation. The main effect is that it allows us to keep track of each

[32] See Stalnaker (1970, 1974, 1978, 1998, 1999: introduction). Note that one can accept Stalnaker's account of conversational context without accepting his views on the rules governing the way contexts change.

others' beliefs and exchange information smoothly. It registers the type of move that is being made in the conversation. When a participant utters a sentence, the force of her utterance informs the other members of the conversation what to expect and how to behave. In David Lewis's terms, force allows us to *keep score* properly in a conversation.[33]

On Lewis's account of conversational scorekeeping, each stage of a conversation is associated with a score. The *score* is a set of abstract entities (he cites presupposed propositions and boundaries between permissible and impermissible courses of action as examples). The score determines both which utterances are acceptable and the contextually determined features of the sentences uttered (e.g., disambiguation, the antecedents of pronouns, the contents of context dependent expressions, the scopes of quantifiers, etc.). The way the score changes from one stage to the next is rule-governed. Lewis's example of a rule that specifies the kinematics of score is: if at time t the conversational score is s, and if between time t and time t′, the course of conversation is c, then at time t′ the score is s′, where s′ is determined in a certain way by s and c.[34] That is, the conversational actions of the participants and the events that occur in their local environment affect the conversational score.

There are several ways to treat the relation between the participants and the rules that specify the kinematics of score. First, the rules might be constitutive in the sense that they define what counts as an acceptable move in the conversation in terms of the behavior of the participant and the score. Second, the score might be operationally defined in the sense that the score is whatever a given scoreboard says it is. Here, the assumption is that the scoreboard is some batch of mental representations. Third, one might define the scoreboard as whatever best fills the role specified by the rules governing the kinematics of score and define the score as whatever the

[33] Lewis (1979a). See Brandom (1994), Lance (1998, 2001), DeRose (2004), and Feldman (2004) for discussion.
[34] Lewis (1979a: 238).

scoreboard says it is (Lewis tentatively accepts this third option).[35] It will not matter for my purposes which way the score is defined.

When a participant in a conversation utters a sentence, that utterance affects the conversational score. The utterance is an attempt at a move in the conversation. Thus, the utterance (qua physical behavior) affects the score in two ways. First, it is an event that occurs in the vicinity of the conversation and is noticed by the participants in the conversation and assumed by the participants in the conversation to be noticed by all the others. Second, it is an attempt at a move in the conversation. That is, it is recorded as an attempt to affect the score with the content of the sentence uttered. Of course, it is intended to have more than these two effects on the conversational score—it is intended to be a legitimate move in the conversation. If the utterance counts as a legitimate move in the conversation, then the effect it has on the score is a product of the force of the utterance and the content of the sentence uttered. Obviously, the force of the utterance is not determined by the syntactic or semantic features of the sentence uttered (e.g., one can use a declarative sentence to ask a question).[36]

If the participants in the conversation cannot determine the force of a particular utterance, then they cannot decide how that utterance alters the conversational score. If they cannot decide how that utterance alters the conversational score, then they cannot alter their own beliefs and expectations about which moves are acceptable. Thus, they cannot determine which future moves are legitimate. In short, if the members of a conversation cannot determine the forces of the utterances made, then they cannot continue with the conversation.

Consider some examples. Assume that Clancy and Sara are having a conversation. Clancy utters ‘everything Ralph uttered in lecture yesterday is true’. Assume that Clancy does

[35] Lewis (1979a: 239-240).
[36] See Davidson (1982) where he argues that if we introduced syntactic or semantic markers for force, then they would be used in most films, plays, and novels as well.

not know which sentences Ralph uttered yesterday in the lecture, but he has good reason to believe that they are true. Thus, Clancy does not know whether the sentence he uttered is paradoxical. Assume that Sara does not know which force Clancy's utterance has. She asks: was that an assertion? Clancy is expected to treat this question as relevant and provide an answer to it. If paradoxical sentences cannot be asserted, then Clancy should respond by saying, "I don't know." The fact that this sort of thing does not happen and would be unacceptable should cast doubt on the doctrine in question.

Nevertheless, let us press on. How should Sara respond to Clancy's admission? Should she attribute a belief to Clancy? Should she expect him to assent to this claim if queried? Should she expect him to act in a way that is consistent with a belief that everything Ralph uttered at yesterday's lecture is true? Is it acceptable for Clancy to appeal to this claim in order to justify another claim later in the conversation? Can Sara challenge Clancy's claim? What sorts of justifications are required to challenge such an utterance? Should Sara adopt the belief that everything Ralph uttered in yesterday's lecture is true? Is it acceptable for her to defer to Clancy if she asserts this claim in another conversation and is challenged to justify it? Is Clancy at fault if it turns out that his utterance is not an assertion? It is unclear how to answer these questions because we do not have a force that is appropriate for utterances that are intended to be assertions but fail.

This example illustrates a further point about conversational score. It is unacceptable in a conversation to perform an utterance without believing that it has a certain force. If asked about the force of one's utterance, one must always be able to specify a force. Does that mean one can never be mistaken about the force of one's utterance? I do not know; but that is not the point here. When attempting to make a move in a conversation, one must intend it to have a certain force and have good reasons for believing that it has that force; otherwise the move is unacceptable. It is always appropriate for one participant to ask another about the force of an utterance and to expect a definitive answer.

Of course, introducing a force for attempted assertions of sentences that turn out to be paradoxical would not help matters at all. The participants of a conversation would not know whether a given utterance is an assertion or whether it has the other type of force. Thus, they would not know how to adjust the score of the conversation appropriately.

Advocates of the claim that utterances of paradoxical sentences are not assertions are fond of citing semantic externalism to support their claims.[37] In particular, the view that there are singular propositions (i.e., propositions some of whose constituents are physical objects) and that some propositional attitudes are relations between people and such propositions implies that if a person mistakenly believes that an object exists, then that person might believe that he has a belief about that object when in fact there is no proposition for him to believe. Thus, what seems to him to be a belief is not a belief at all.[38] The analogy to the view of assertion I have been discussing is that if one mistakenly believes that a sentence is non-paradoxical, then that person might believe that he has asserted it, when in fact his utterance does not count as an assertion. This analogy is especially tight for those who treat speech acts on the model of propositional attitudes (i.e., John's assertion that p is a relation between John and the proposition that p). If paradoxical sentences do not express propositions, then there is no proposition for John to assert.

There are several problems with an appeal to this version of semantic externalism to justify the claim that utterances of paradoxical sentences are not assertions. First, appealing to this version of semantic externalism is not a good way to garner support for one's view given

[37] See Goldstein (2001). Glanzberg draws a similar analogy to support his claim that utterances of sentences that are truth-value gaps are not assertions; see Glanzberg (2003).
[38] See McDowell (1977, 1984) and Evans (1982).

that this version of semantic externalism is rather controversial. Second, there are good reasons (or at least plausible reasons) for thinking that there are singular propositions and that if the object in question does not exist, then the purported belief is not a genuine belief. There is no analogous argument for the claim that utterances of paradoxical sentences do not count as assertions. Thus, there is no good reason to think that an utterance of a paradoxical sentence is not an assertion other than that it seems to some to help deal with the liar paradox.

A defender of the claim that paradoxical sentences cannot be asserted might object that the same reasoning that leads one to claim that one cannot have certain beliefs about an object if that object does not exist supports the claim that utterances of truth attributions that turn out to be paradoxical are not assertions. All paradoxical sentences attribute some semantic property (e.g., truth, falsity, gaphood, etc.) to some object or objects. That is not quite right; rather, paradoxical sentences purport to attribute some semantic property. In particular, they purport to attribute some semantic property to a proposition or propositions. Usually, the purported target of the attribution is what is taken to be the proposition expressed by the very sentence itself. Consider λ (the liar sentence token). λ purports to attribute falsity to itself or rather, to the proposition it supposedly expresses. However, paradoxical sentences do not express propositions. Thus, the target of the attribution does not exist. Thus, an utterance of λ would not count as an assertion for the same reason that a purported belief about an object that does not exist does not count as a belief—in both cases, the target does not exist.

This reply to my objection depends on the claim that paradoxical sentences do not express propositions. In the next section, I show that this view is unacceptable as well, and for essentially the same reasons.

B.6 Propositions

One of the perennial issues in the literature on truth is the choice of truth bearers. Sentence tokens, sentence types, propositions, statements, beliefs, utterances, assertions, and computational roles have all been suggested as truth bearers. No one disputes the fact that it is common to attribute truth to all these things, and no one thinks that these are unrelated properties that coincidentally have the same name. They are intimately connected. The most common way of explaining their connection is to provide an account of truth for primary truth bearers and extend the account to other types of truth bearers. There seem to be two ways to go on this issue. One is to say that there is only one type of truth bearer, and any time someone attributes truth to a different type of entity, that is just shorthand for a truth attribution to a real truth bearer. For example, if one accepts that propositions are the primary truth bearers, then one treats an attribution of truth to a sentence token as an attribution of truth to the proposition expressed by that sentence token. The other strategy is to designate one type of entity as primary truth bearer, give an account of truth for them, and then extend it to other types of entities by their relations to the primary truth bearers. For example, if propositions are the primary truth bearers and one has an account of truth for propositions, then one can provide an account of truth for sentence tokens by saying that a sentence token is true if and only if it expresses a true proposition. Combinations are also possible; e.g., propositions are primary truth bearers and a sentence token is true if it expresses a true proposition, but attributing truth to a sentence type is just shorthand for saying that a token of that type is true (because sentences types are not truth bearers).

Propositions are one of the most popular candidates for truth bearers.[39] There are several contemporary theories of propositions, but they fall into two broad categories. The first is that propositions should be explained in terms of possible worlds. The standard account of this sort is that a proposition is a function from a set of possible worlds to a set of truth-values. The second is that propositions are structured entities. That is, a proposition has constituents that are usually taken to be the semantic values of the linguistic expressions occurring in a sentence that expresses that proposition.[40] It does not matter for my purposes which account of propositions one takes to be the correct one. I do want to point out that propositions are intended to serve several purposes in philosophical theories. First, they are intended to be the primary bearers of truth and falsity. That role is my focus in this section. Second, they are supposed to serve as the objects of propositional attitudes (e.g., belief, desire, intention). That is, when someone has the belief that p, there is a relation between that person and the proposition that p. Third, it is assumed by defenders of propositions that they are the contents of sentence tokens on occasions of use. That is, when a person utters the sentence token q in a conversational context, q expresses a particular proposition (or multiple propositions). The proposition expressed by a sentence token in the context in which it is uttered encapsulates the truth conditions conveyed by the sentence token in that context. Finally, propositions are sometimes used as the bearers of epistemological and modal properties (i.e., knowledge, justification, necessity, possibility, etc.).

I argue that the riskiness thesis implies that propositions are a poor choice for primary truth bearers. Here is the problem. Either paradoxical sentences express propositions or they do not. If paradoxical sentences do not express propositions, then whether a sentence token uttered in a conversation expresses a proposition is not available to the participants of that conversation

[39] See Barwise and Etchemendy (1987), Horwich (1998), Soames (1999), Glanzberg (2001, 2003, 2004), and Künne (2003) for examples.

[40] See Stalnaker (1987) for an example of the former and King (1995) and Soames (2002) for examples of the latter.

(because of the riskiness thesis). I use a version of the availability argument to show that this result is unacceptable. If paradoxical sentences do express propositions, then one can prove that certain sentence tokens both express and fail to express propositions. That is, one can generate a new paradox.

To see that the claim that paradoxical sentences do not express propositions is unacceptable, consider another version of the availability argument:

(1P) Paradoxicality need not be available to the participants in a conversation.

(2P) Propositional expression (i.e., whether a sentence token uttered expresses a proposition) is available to participants in a conversation.

∴ (3P) Propositional expression does not depend on paradoxicality (i.e., whether a sentence token uttered expresses a proposition does not depend on whether the sentence token is paradoxical in context of utterance).

I argued for (1P) in the previous section (under the name '(1A)'). The argument for (2P) is similar to the argument for (2A). If the participants of a conversation could not determine whether a given sentence uttered expresses a proposition, then they would not be able to keep score in a way that would allow the conversation to continue. Both the force of an utterance and the proposition expressed by the sentence token uttered must be available to the participants in a conversation for their interactions to constitute a conversation.

There is good reason to claim that paradoxical sentences do express propositions. However, the claim that paradoxical sentences express propositions encounters troubles of its own. The following is an argument that is well known to those who work on the liar paradox, but the particular formulation of it I present is due to Glanzberg.[41] If we symbolize propositional expression by E(x, y) (i.e., sentence token x expresses proposition y), and propositional truth in a

[41] Glanzberg (2004: 33-34).

possible world by T(x) (i.e., proposition x is true), then we can formulate a propositional liar sentence:

$(\Pi) \sim (\exists x)\, E\, (\pi, x) \wedge T(x)$

If 'π' is the name of the sentence token of Π, then we can prove a contradiction from the claim that π expresses a proposition and several auxiliary claims about propositional expression. The additional claims are:

(E1) $E(\langle s \rangle, p) \supset (T(p) \equiv s)$ (i.e., if a given sentence expresses a proposition then the sentence that attributes truth to that proposition is materially equivalent to the sentence that expresses that proposition).

(E2) $(E(p, q) \wedge E(p, r)) \supset q = r$ (i.e., a sentence token expresses a unique proposition).

(E3) $p = q \supset (T(p) \equiv T(q))$ (i.e., truth is an extensional property of propositions).

From the claim that λ expresses a proposition (i.e., $(\exists x)(E(\lambda, x))$ and principles (E1) – (E3), we can derive a contradiction.[42] That is, we can prove that if (E1) – (E3) are true, then λ does not express a proposition.

One might object that although I have shown that some paradoxical sentences cannot express propositions, I have not shown that some empirically paradoxical sentences cannot express propositions. I agree. Fortunately, it is not difficult to construct an empirically paradoxical sentence that cannot express a proposition. Consider a variant of the example in section two in which α and β are the propositional variants of the sentences in the old example. That is, α is a token of '$(\exists x)(E(\beta, x) \wedge T(x))$', β is a token of '$(\exists x)(E(\gamma_3, x) \wedge T(x))$' and γ_3 is a token of the following type:

[42] Assume $(\exists x)E(\lambda, x)$. Let 'a' be the name of the proposition λ expresses. Assume T(a). Thus, $\sim (\exists x)(E(\lambda, x) \wedge T(x))$ (by (E1)). Therefore, $\sim T(a)$. Assume $\sim T(a)$. Thus, $(\exists x)(E(\lambda, x) \wedge T(x))$ (by (E1)). Therefore, T(a) (by (E2) and (E3)). Consequently, $\sim T(a) \equiv T(a)$.

> Γ_3: Either the proposition expressed by β is false or there are more objects heavier than one gram above the ecliptic of the Earth's solar system than there are objects heavier than one gram below the ecliptic at 1200 GMT on January 1, 2004.

(The ecliptic is the plane that contains the closed curve marked out by the Earth's orbit; 'above' is defined in the obvious way via the positive orientation of the orbit curve.) In both worlds (i.e., w_1 and w_2), Stu utters α, β is inscribed on the blackboard in room 1, and γ_3 is inscribed on the blackboard in room 2. The only difference between w_1 and w_2 is the distribution of matter in our solar system such that that the second disjunct of γ_3 is false in w_1 and true in w_2. We can show that in w_1, α, β, and γ_3 are paradoxical, while in w_2, α, β, and γ_3 are true. Let 'δ' be the name of the sentence token that occurs as the second disjunct of γ_3. Assume that is w_1, δ is false and in w_2, δ is true.[43] In w_1, α is paradoxical, while in w_2, α is true. Here is the argument. Consider w_1. Assume β is true. If β is true, then γ_3 is true. If γ_3 is true, then either δ is true or β is false. δ is false. Thus, β is false. Hence, if β is true, then β is false. Assume β is false. If β is false, then γ_3 is false. If γ_3 is false, then δ is false and β is true. Thus, β is true. Hence, if β is false, then β is true. Therefore, β is true if and only if β is false. Consequently, β is paradoxical. It follows that in w_1, α, β and γ_3 are paradoxical. However, in w_2, γ_3 is true (because δ is true), which implies that β and α true as well.

I have shown that both the claim that paradoxical sentence tokens express propositions and the claim that paradoxical sentences do not express propositions are unacceptable. Thus, propositions are a poor choice for primary truth bearers.

[43] Of course, the truth-value of δ will depend on how we define 'object', 'gram', 'ecliptic' and 'solar system'. As the example is formulated, I am assuming that these terms would be given definitions that do not display indeterminacy. The example can, of course, be modified to accommodate indeterminacy.

B.7 Minimalist Theories of Truth-Aptness

Although it might be tempting to think of truth-aptness issues and truth bearer issues as one and the same, it seems to me that there is an important distinction here. The issue of truth-aptness arises once one has already made choices about truth bearers. A truth bearer is truth-apt if it is capable of having the property of truth.[44] How is this any different from just being a truth bearer? The difference is that one's choice of truth bearers is a choice between types of objects without regard to their semantic or pragmatic features. The options for truth bearers are not distinguished in semantic terms (e.g., fact-stating sentence tokens). If one chooses sentence tokens as truth bearers, then there is still the issue of deciding which sentence tokens are capable of possessing truth. We do not think that sentence tokens used to produce questions or commands are true or false. Some philosophers argue that sentence tokens whose semantic presuppositions fail are neither true nor false.[45] Others claim that sentence tokens that contain occurrences of normative vocabulary are neither true nor false.[46] These are truth-aptness issues. A single truth bearer can be truth-apt in one context and not truth-apt in another. For example, 'the room is cool'. If a token of this sentence is used to describe the temperature of the room in question then it is truth-apt (on most accounts), but if it is used to express one's positive evaluation of the room, then, on some views, it is not truth-apt.

Of course, one could combine truth-aptness issues and truth bearer issues into one topic, but I think that this would do a disservice to those engaged in debates about them. Expressivists

[44] I have heard some people use the term 'truth-aptitude' instead of 'truth-aptness'. As I understand the terms, 'aptness' and 'aptitude' have very similar meanings, but the latter tends to have the connotation of ability—something that an animate entity can do—whereas the former seems to apply more readily to inanimate objects. I prefer 'aptness' since it seems odd to me to say that truth-apt truth bearers have the ability to be true (i.e., they can accomplish truth if they try hard enough).

[45] See Strawson (1952).

[46] See Gibbard (1990) for an example.

who discuss which things are capable of truth are not (usually) worried about choosing between propositions and sentence tokens; they are concerned with truth-aptness. Another way to bring out the difference is to say that truth bearer issues are explanatory (i.e., how should we go about explaining the truth and falsity of one type of entity in terms of the truth and falsity of another?), whereas truth-aptness issues are demarcational (i.e., which entities are in the class of those that can be true or false?). Clearly, these are different issues and deserve to be kept distinct.

For clarity, I distinguish truth-valuedness from truth-aptness and truth bearerhood. *Truth-valuedness* is the property of being true or false. On some accounts, a truth-apt truth bearer can fail to be truth-valued. The distinction between truth-aptness and truth-valuedness is intended to allow for the view that some truth bearers are truth-apt but, because they are paradoxical, they are not truth-valued (i.e., they are gappy). There are important differences between a failure to be truth-apt and a failure to be truth-valued. One is that if a person attributes truth to a sentence token (acceptable truth bearer) that is used to produce a question (not truth-apt), then that person does not have a full grasp of the concept of truth. In some sense, he has a made a category mistake. However, if a person attributes truth to a sentence token (acceptable truth bearer) that is truth-apt, but paradoxical (not truth-valued on some accounts), then that person is probably not at fault for attempting to apply truth to an object that is not true or false. Language users are often not at fault for attributing truth or falsity to paradoxical sentences because the fact that they are paradoxical is beyond what any responsible language user could be expected to know. Such a person has not made a category mistake and might well have a perfect grasp of the concept of truth. Truth bearerhood and truth-aptness are properties that should be available to language users, while truth-valuedness might not be. We treat the failure to be a truth bearer, failure to be truth-apt, and the failure to be truth-valued rather differently.

Distinguishing between attributions of truth to objects that are not truth bearers, attributions of truth to truth bearers that are not truth-apt, and attributions of truth to truth-apt truth bearers that are not truth-valued is essential to understanding the issues that arise in connection with the liar paradox.

The notion of truth-aptness is a recent addition to discussions of the nature of truth and it comes up almost exclusively in connection with the compatibility of deflationism and nonfactualism. Nonfactualism about X is the view that X talk is not representational, that it does not purport to describe the world, or that there are no facts about Xs. Often, nonfactualism about X is thought of as the claim that sentences involving X are not truth-apt. There is an important debate about whether deflationism implies that all declarative sentences are truth-apt, which has led to a distinction between minimalist and substantive accounts of truth-aptness.[47] The two minimalist accounts of truth-aptness are syntacticism and disciplined syntacticism.

Syntacticism is the view that truth-aptness depends only on the syntactic features of a truth bearer. Obviously, syntacticism is compatible only with theories that take truth bearers to have syntactic properties. The most common syntactic property cited by syntacticists is that of being a declarative sentence. The most common criticisms of syntacticism are that there are declarative sentences that are not truth-apt and that it takes more than syntactic properties to insure that a truth bearer is truth-apt.[48]

Disciplined syntacticism is the theory that a truth bearer is truth-apt if and only if it has the right syntactic features and interacts properly with other common linguistic expressions (i.e., it can be embedded in truth functional contexts, 'S believes that' can be appended to it, etc.).

[47] See Boghossian (1990), Kraut (1993), Dreier (1994), Smith (1994), Divers and Miller (1994), Horwich (1994), Smith (1994), Kalderon (1997), Wedgwood (1997), Blackburn (1998), Wright (1998), Holton (2000), Swan (2000), Jackson et al. (1994), Smith (1994), Holton (2000), Dodd (2002), and Engel (2002).

[48] Wright (1992); see Jackson et al. (1994) for discussion.

Criticisms similar to those that apply to syntacticism can be wielded against disciplined syntacticism as well. Usually theories that require more for truth-aptness than disciplined syntacticism does are called *substantive*.[49]

Several deflationists have endorsed substantive accounts of truth-aptness. There is very little common ground here so making generalizations is difficult, but some substantive theories of truth-aptness require truth-apt sentences to be capable of expressing beliefs; others demand that the presuppositions of the sentences in question are met.[50] However, many philosophers who discuss deflationism assume that it is incompatible with accounts of truth-aptness that are more substantive than disciplined syntacticism.[51]

Enough history. If one accepts either syntacticism or disciplined syntacticism, then α in w_1 and α in w_2 have to have the same truth-status (where a sentence token's truth status is either its truth-value or its lack of truth-value). However, α in w_1 is paradoxical and, hence, not true, but α in w_2 is true. Thus, α in w_1 and α in w_2 have different truth statuses. Therefore, syntacticism and disciplined syntacticism are false.

If deflationists have to accept either syntacticism or disciplined syntacticism then deflationism is sunk by this argument as well. However, I see no reason to think that deflationists are required to accept a minimalist theory of truth-aptness.

If one wanted to use syntacticism or disciplined syntacticism as a theory of truth-valuedness instead of as theory of truth-aptness, then we encounter what is essentially the same problem again. Truth-valued sentence tokens are not paradoxical. However, paradoxicality does

[49] Wright (1992).
[50] See Kraut (1993) for an example.
[51] See the debate outlined in fn. 47.

not depend on the syntactic features of a sentence token. Thus, neither syntacticism nor disciplined syntacticism is capable of serving as a theory of truth-valuedness.

B.8 T-SCHEMA DEFLATIONISM

One of the most popular versions of deflationism is *T-schema deflationism*: a theory of truth-in-L for a language L consists of all and only the T-sentences for the sentences of L.[52] A *T-sentence* is a sentence of the form: ⟨p⟩ is true if and only if p. If the language in question is classical, then the connective in the T-sentences is the material biconditional.[53] If the language in question is non-classical, then there is a range of options. One can use a weak Kleene or a strong Kleene biconditional (both of which are monotonic and value-theoretic), a Łukasiewicz or a Holton biconditional (both of which are non-monotonic but value-theoretic), or any one of the many non-monotonic intensional biconditionals on the market.[54]

One obvious problem for T-schema deflationism is that the set of T-sentences for most any language that has minimal expressive capacities is inconsistent (in classical logic) because of the liar paradox and its brethren. The three main responses to this problem are: (i) restrict the set of T-sentences that constitute the theory to those for non-paradoxical sentences of the language, (ii) weaken the logic and keep all the T-sentences, and (iii) use a different biconditional and keep

[52] Truth-in-L is a language-specific concept of truth (an *LS* concept). The extension of 'true-in-L' is the set of true sentences of L (one can either treat it as a partial predicate and claim that its anti-extension is the set of false sentences of L, or one can treat it as a completely-defined predicate and claim that its extension and anti-extension are jointly exhaustive). T-schema deflationism is a theory of truth-in-L for a particular language L. An advocate of this theory might claim that natural language truth predicates can be explained in terms of LS truth predicates. See Appendix A for a criticism of explaining natural language truth predicates in terms of LS truth predicates.

[53] Of course, many deflationists consider the T-sentences to be necessary; hence, they treat the connective as strict co-implication.

[54] The most promising of these is Field's biconditional; see Field (2002, 2003a, 2003b, 2003c, 2004, forthcoming a, forthcoming b, forthcoming c). See also Gupta and Belnap (1993), Beall (2000), and Yablo (2003).

all the T-sentences (combinations of (ii) and (iii) are also promising). Accordingly, there are two types of T-schema deflationism: those that exclude T-sentences for paradoxical sentences from the theory (option (i)) and those that include all the T-sentences (options (ii) or (iii)). I call the former *exclusive* and the latter *inclusive*.[55] The target of my criticism is exclusive T-schema deflationism. For my purposes, it makes no difference which biconditional is used for non-classical languages by the theories in question, so long as T-sentences for paradoxical sentence tokens are excluded from the theory of truth-in-L.

McGee dealt a blow to exclusive T-schema deflationism (from here on I suppress the 'exclusive') by proving that for any set S of sentences of a language L that is consistent with basic facts about the syntax of L, there is a maximally consistent set of T-sentences that is consistent with the basic facts about the syntax of L and that entails S. Thus, a deflationist cannot simply say that she wants a maximally consistent set of T-sentences for her theory because there are lots of them and they are incompatible with one another. Moreover, the overlap between them contains only truth-tellers (sentences like 'this sentence is true' that affirm their own truth). Finally, none of the maximally consistent sets of T-sentences is recursively axiomatizable. Thus, there is no effective way of constructing a maximally consistent set of T-sentences.[56]

McGee's results on maximally consistent sets of T-sentences certainly close off one avenue for a deflationist reply to the semantic paradoxes. Namely, a deflationist cannot simply stipulate that a maximally consistent set of T-sentences provides an implicit definition of 'true'. There are too many of them, they are incompatible, they are not recursively axiomatizable, and

[55] See Horwich (1998) for a version of the exclusive type (Horwich formulates his theory in terms of propositions but claims that it is just as plausible when formulated for sentence tokens). See Gupta and Belnap (1993) and the Field references in the previous footnote for examples of inclusive theories.

[56] McGee (1992). See also Weir (1996) and Gauker (2001).

they overlap only on pathological sentences. Thus, if a deflationist wants to use the T-sentences as implicit definitions of 'true' and wants a consistent definition, then she should provide some way of choosing between them. The received view is that an approach to the paradoxes that characterizes paradoxicality can be use to determine which T-sentences should be included in the theory of truth for the language in question (e.g., McGee suggests the revision theory).[57]

Given the riskiness thesis (i.e., the syntactic and semantic features of a sentence token need not determine whether it is paradoxical), T-schema deflationism is a non-starter. Consider the T-schema theory of truth for the idiolect of English (call it *L*) in the example from section two. Let T_1 be the set of T-sentences for sentences of L that are non-paradoxical in w_1 and T_2 be the set of T-sentences for sentences of L that are non-paradoxical in w_2. The T-sentence for α, ''the sentence written on the blackboard is true-in-L' is true-in-L if and only if the sentence written on the blackboard is true-in-L', is a member of T_2, but it is not a member of T_1. Because T-schema deflationists take the set of T-sentences in question to be definitional of the concept of truth, the concept of truth-in-L in w_1 is different from the concept of truth-in-L in w_2. Call the former *truth$_1$-in-L* and the latter *truth$_2$-in-L*. The T-sentence for α is constitutive for truth$_2$-in-L but not for truth$_1$-in-L.

Is this result plausible? Consider the variant of the example in which γ_3 is the sentence inscribed on the blackboard in room 2. In w_2, the concept of truth-in-L is truth$_2$-in-L and not truth$_1$-in-L because of the particular distribution of matter in our solar system at noon GMT on January 1, 2004. If that distribution had been different, then the concept of truth for L would have been different. That is highly counterintuitive. That is, the concept of truth-in-L does not depend on the distribution of matter in our solar system.

[57] McGee (1992).

One possible reply to this objection is that truth predicates are context dependent—their extensions change from context to context. Thus, contrary to the conventional wisdom, deflationism is not only compatible with a contextual approach to the paradoxes, one version of deflationism implies that truth is context-dependent.[58] In the next section, I pose a criticism for contextual approaches to the paradoxes. After that, the consequence of context dependence should seem much less palatable.

B.9 Contextual Approaches to the Liar Paradox

In section five, I argued that the riskiness thesis implies that it is unacceptable to claim that utterances of paradoxical sentence tokens are not assertions. One objection to my argument depends on the claim that paradoxical sentence tokens do not express propositions. In section six, I argued that it is unacceptable to claim that paradoxical sentence tokens do not express propositions. I also argued that the riskiness thesis implies that propositions are a poor choice for primary truth bearers. An objection to my argument depends on the claim that truth displays an element of context dependence. Furthermore, in section eight, I argued that exclusive T-schema deflationism is unacceptable. One objection to my claim is that truth displays an element of context dependence. Thus, several of my arguments depend on a rejection of contextual theories of truth. In this section, I present several objections to such theories.

There are at least a dozen contextual theories of truth, so instead of either describing each one in detail or picking one as a representative, I address two main types of them. Theories of the first type imply that ‘true’ is a context dependent expression similar to demonstratives (e.g.,

[58] See Simmons (1999) for an argument that contextualist theories of truth and deflationist theories of truth are incompatible.

‘that’), indexicals (e.g., ‘here’), and graded adjectives (e.g., ‘tall’). On a theory of the first type, ‘true’ has a fixed meaning, but its content on an occurrence of its use depends on the context in which it is used. That is, the contribution an occurrence of ‘true’ makes to the truth conditions of a sentence token in which it occurs depends on the context in which that sentence token is uttered. Once the meanings of the subsentential components, the referents of the singular terms, and the logical form of the sentence token have been determined, any ambiguities have been resolved, and the context-determined contents of any non-semantic terms have been determined, there is the additional variability in the semantic content of the sentence token due to the presence of the truth predicate. In particular, the extension of the truth predicate changes from context to context. What would count as a paradoxical sentence token in a given context is eliminated from the extension of ‘true’ in that sentence token. Paradoxical sentence tokens are treated as either false or gappy in the contexts in question.[59]

The theories of the first type solve the liar paradox by finding a context shift in the associated argument. Consider a token λ of ‘λ is not true’ in a context C. In context C, neither the extension nor the antiextension of ‘true’ contains λ. Thus, in context C, λ is a gap. However, if we assert ‘λ is not true’, then we have changed the context to C′. In C′, λ is in the antiextension of ‘true’; hence, ‘λ is not true’ is true in C′, but the token of the same type (i.e., λ) in context C is a gap.

The second type of contextual theory of truth implies that, while ‘true’ is not a directly context dependent expression, sentence tokens in which ‘true’ occurs do display context

[59] Thomason (1976), Burge (1979a, 1982a, 1982b), Gaifman (1982, 1992, 2000), Hodges (1986), Barwise and Etchemendy (1987), Koons (1992, 2000b), Simmons (1993, 2000, 2003), and Cantini (1995). For criticism of Burge, see Gupta (1982) and Simmons (1993); for criticism of Gaifman see Simmons (1993) and Yi (1999); for criticism of Barwise and Etchemendy, see Gupta (1989), Grim (1991), McGee (1991), Gaifman (1992), and Priest (1993); for criticism of Simmons, see Antonelli (1996), Hardy (1997), and Beall (2003). For discussion of revenge liars for contextual theories see Hazen (1987), Hinckfuss (1991), Juhl (1997), Clark (1999), Weir (2000, 2001), and Weir (2002).

dependence. One way to think of this context dependence is to treat truth as a predicate of propositions and claim that an attribution of truth to a sentence is actually an attribution of truth to the proposition expressed by that sentence. On this view, attributions of truth to sentences have a hidden quantifier. For example, let ‘p’ be the name of a sentence token. The truth attribution ‘p is true’ becomes ‘the proposition expressed by p is true’; on a Russellian interpretation of definite descriptions, the latter becomes ‘there is a proposition such that it is expressed by p and it is true’. Quantifiers are known to display context dependence in their scope. That is, the set of objects over which a quantifier ranges is determined, in part, by the context in which it is used. It is this quantificational context dependence that is claimed to be present in sentences that contain truth predicates. The appeal to propositions can be replaced with talk of schemes for interpreting sentences, which contain domains and are determined by the context.[60]

Theories of this type solve the liar paradox by claiming that one sentence token does not express a proposition at all, but another token of the same type does express a proposition. For example, if λ is a token of ‘λ is not true’, then λ actually says that it is not the case that there exists a proposition such that it is expressed by λ and it is true. Assume that λ is in context C. In context C, the scope of the hidden quantifier in λ is a certain set of propositions. In C, there is no proposition in this set for λ to express. However, if one asserts ‘λ is not true’ then this action changes the context. In the new context C′, the quantifier ranges over a more inclusive set of propositions, one of which is the proposition that it is not the case that there exists a proposition such that it is expressed by λ and it is true.

[60] Parsons (1974, 1983, 1984) and Glanzberg (2001, 2004).

Most contextual theories of truth (of either type) incorporate some sort of hierarchy in order to accommodate changes in context. For example, Burge's theory employs a Tarskian hierarchy of truth predicates as a supply of contents for 'true'. In a given context, 'true' has one of the contents of a Tarskian truth predicate. Other examples include Parsons, who employs a hierarchy of interpretation schemes, and Glanzberg, who employs a hierarchy of contexts. Although Simmons' theory does not employ a hierarchy of contexts, it does appeal to a hierarchy of invariant truth predicates to handle revenge paradoxes.

I make three points in the remainder of this section: (i) given the riskiness thesis, the standard arguments used to support contextualist theories of truth are circular, (ii) the riskiness thesis casts doubt on whether contextual theories of truth can respect a Gricean condition on theories of context, (iii) the riskiness thesis implies that if we adopted a context dependent truth predicate that behaves in the way contextualists claim, then it would be an impediment to communication.

One of the driving forces behind the contextual approach is that it handles cases like:

λ_1: λ_1 is not true.

λ_2: λ_1 is not true.

'λ_1' and 'λ_2' are to be thought of as names of the particular physical sentence tokens on this page of this particular physical document.[61] λ_1 and λ_2 are tokens of the same type. However, on one interpretation, while λ_1 is a liar sentence and thus, paradoxical, λ_2 is a comment on λ_1 to the effect that, because it is paradoxical, it is not true. Thus, on this interpretation, although they are

[61] That stipulation makes the tokens 'λ_1' and 'λ_2' ambiguous because there are multiple copies of this document. This complication does not affect my discussion.

two tokens of the same type, one is true and the other is paradoxical. This phenomenon is known as *the two-line puzzle*.[62]

A related phenomenon concerns the reasoning that accompanies λ_1. An intuitive argument supports the claim that sentence λ_1 is true if and only if it is not true. Thus, sentence λ_1 seems to be paradoxical. It is natural to think that paradoxical sentences are not true because assuming that they are true leads to contradiction. Thus, one might conclude that because sentence λ_1 is paradoxical, it is not true (indeed, I have argued this point above). At this point, one might reread sentence λ_1 and realize that it says of itself that it is not true. We have just argued that it is not true; thus, sentence λ_1 accurately describes its own status—it says that it is not true and, indeed, it is not true. Hence, sentence λ_1 must be true (for that is what we say about sentences that say that such and such is the case when such and such actually is the case). I will refer to this as *the strong liar reasoning*.[63]

The first point is that because paradoxicality affects the truth status of a sentence token and paradoxicality is not determined by the syntactic and semantic features of a sentence token, it is inappropriate to argue from the claim that two sentence tokens of the same type differ in truth-status to the claim that they must contain a context-dependent expression. The assumption at work in the argument for contextual theories is that if two sentence tokens of the same type have different truth-statuses (as is the case with λ_1 and λ_2), then they must contain a context-dependent expression. I call this the *context-dependence principle*. The argument in the case at hand begins with the assumption that λ_1 is paradoxical and λ_2 is true. Thus, they have different

[62] See Hazen (1987), Gaifman (1992), Juhl (1997), Clark (1999), Goldstein (1999, 2001), Weir (2000, 2002), and Gupta (2001).

[63] See Kearns (1970), Parsons (1974), Burge (1979), Hazen (1987), Gaifman (1992, 2000), Gupta (2001), and Glanzberg (2001, 2003, 2004).

truth statuses but are tokens of the same sentence type. Hence, they contain a context-dependent element.

The problem with this argument is that if paradoxicality is not determined by the syntactic and semantic features of a sentence token and paradoxical sentences are not true, then the principle used above is false. As a variant of the example in section two shows, one construct two sentence tokens of the same type, one of which is paradoxical and the other of which is true, yet both have the same syntactic and semantic features. In order to appeal to the context-dependence principle, the contextualist must reject either the riskiness thesis or the assumption that paradoxical sentences are not true. However, only contextual theories of truth reject either of these claims.[64] Thus, in appealing to the context-dependence principle to justify a contextual theory of truth, a contextualist is appealing to a principle that only a contextualist would endorse. Analogous criticism applies to contextualists who appeal to the strengthened liar reasoning to justify the claim that truth predicates are context-dependent.

My second point is that there is a tension between contextual theories of truth and the Gricean intuition that a central goal of linguistic activity is communication. Stalnaker expresses one consequence of this intuition in the following passage:

> It is a substantive claim that the information relevant to determining the content of context-dependent speech acts is presumed to be available to the participants of a conversation—that it is included in the presuppositions of the context—but it is a claim that is motivated by natural assumptions about the kind of action one performs in speaking. It is not unreasonable to suppose that speakers, in speaking, are normally aiming to communicate—at least to have the addressees understand what is being said. Succeeding in this aim requires that the information relevant to determining content be available to the addressee. The representation of context as a body of presupposed information is also appropriate to the other side of the interaction between context and content, since it is reasonable to suppose that a body of information is also what speech acts act on. If the goal of speech, or at least one central goal, is to exchange information, then

[64] As I have said, others reject the riskiness thesis as well but these views are implausible on their own and incompatible with contextualism; e.g., dialetheism accepts that paradoxical sentences are both true and false.

> it is natural to explain the force of speech acts as the attempt to add to or alter a body of information that is presumed to be shared by the participants in the conversation, (Stalnaker 1999: 6).

Stalnaker argues that the aspects of the conversational context that determine the content of the sentence tokens uttered in that context should be available to the participants in that conversation. If not, then the audience members cannot use the relevant aspects of the context to determine the content of the sentence token uttered and the speaker cannot use the relevant aspects of the context to determine the potential contents of sentence tokens he intends to utter. Stalnaker's theory of conversational contexts respects this condition.

The example in section three shows that if contextualist theories of truth are correct that the content of a sentence token containing an occurrence of a truth predicate depends on which sentence tokens are paradoxical, then the theory of conversational context required by the contextualist theory of truth will not satisfy the Gricean condition voiced by Stalnaker in the above passage. In other words, if contextualist theories of truth are correct, then they cannot rely on Stalnaker's theory of conversational context or any other theory that implies that the aspects of the conversational context are available to the participants in the conversation.

This objection directly affects Simmons who advocates both a contextualist theory of truth and Stalnaker's theory of conversational context.[65] However, the objection has wider implications. A proponent who endorses a contextualist theory of truth must adopt or construct a theory of conversational context on which the relevant aspects of the context need not be available to the participants in the conversation. Furthermore, a philosopher in this position should explain how utterances of sentence tokens containing occurrences of truth predicates could be used in a process of successful communication. If the participants in a conversation do not have the resources to determine the content of sentence tokens uttered in the conversation,

[65] See Simmons (1993, 2000) for the theory of truth and Simmons (2003) for the endorsement of Stalnaker's theory.

how can they communicate successfully? If speakers are unable to determine what the content of a sentence token in which a context dependent truth predicate occurs would be if uttered and audience members are unable to determine the content of such sentence tokens when uttered, then why would anyone employ a context dependent truth predicate or participate in a conversation in which one is used? Without answers to these questions, contextualist theories of truth are unable to explain the most obvious fact about our practice of using truth predicates—the fact that we continue to use truth predicates at all.

My third and final point is that if a content-determining feature of a conversational context is not available to the participants of a conversation, then it must be *relatively constant* in the sense that it is either present in the vast majority of the conversations or absent in the vast majority of the conversations in which the expression in question is employed. Otherwise, the participants in the conversation would have no way of successfully communicating. If the feature in question is relatively constant, then the participants in the conversation can be justified in assuming that the feature in question is present (absent) for almost all the utterances made in a conversation. That is, they can still successfully communicate by "factoring out" the feature in question.

For example, assume that those who live in the town of Varyville employ an expression whose content depends on the number of rocks on the mantle of the mayor's fireplace. These people use the term 'in-town', which is a context dependent term. If there are no rocks on the mayor's mantle, then the extension of 'in-town' is the area within ten meters of the fountain in the town square. If there is one rock on the mayor's mantle, then the extension of 'in-town' is the area within one hundred meters of the fountain in the town square. If there are two or more rocks on the mayor's mantle, then the extension of 'in-town' is the area within one kilometer of the fountain in the town square. Thus, ‘in-town’ is a context dependent expression. Its content depends on the context of the conversation in which it is employed. It just so happens that Varyville has had one mayor for the past 50 years and he has had a single rock on his mantel for the entire time. Thus, although ‘in-town’ is a context dependent expression, every time it has been uttered in the past 50 years, it has had the same content. Thus, although the number of rocks on the mayor’s mantle is not available in most conversations in which the citizens of Varyville participate, they can use ‘in-town’ and assume that they mean the same thing when they use it. Because the content-determining features of ‘in-town’ do not change during the course of a conversation or from conversation to conversation, the members of Varyville know that their uses of ‘in-town’ have the same content.

Despite the fact that the content-determining features of the conversational context for ‘in-town’ are often unavailable to the participants of conversations, the citizens of Varyville can employ it without running into trouble because the content-determining feature is relatively constant. Contrast the story about Varyville with the account of truth given by the advocate of a contextualist theory of truth. On the contextualist’s theory, the content of a sentence token containing a truth predicate is determined by the context in which it is used. However, the context in which the sentence token is used might have to be very broad. In the example given in section six, the context includes the distribution of matter in our solar system. Of course, one could construct a sentence token whose paradoxicality depends on most any fact. Thus, there is no limit to what might count as a content-determining feature of a given conversational context. Furthermore, the content-determining features are not relatively constant. Indeed, they change continuously. Given any empirical fact, one can construct a sentence token whose paradoxicality depends on that empirical fact. Because empirical facts change continuously, the set of sentence tokens that count as paradoxical changes continuously. Hence, the content-determining features for contextual truth predicates change continuously. Therefore, users of context dependent truth predicates have no reason to think that they mean the same thing in a conversation.

Similar criticisms have been leveled against semantic externalism as well. Semantic externalism is, roughly, the doctrine that the content of some linguistic expressions (and some mental states) depends on the physical or social environment of the users of those expressions (and the possessors of those states).[66] I assume that the reader has a superficial familiarity with semantic externalism. An objection to my criticism of contextual theories of truth is that defenders of semantic externalism have argued in the face of similar criticisms that there are plausible accounts of semantic knowledge (a person's knowledge of the contents of the linguistic expressions he uses and the mental states he possesses) that are compatible with both semantic externalism and the claim that competent language users have semantic knowledge (and even that this knowledge is privileged in certain ways).[67] An advocate of a contextual theory of truth might suggest that the same accounts of semantic knowledge would be compatible with a contextual theory of truth.

The problem I posed for contextual theories of truth is different from the problem semantic externalists face. I point out two differences. First, in the case of semantic externalism, there are good reasons (or at least there are some arguments that some philosophers find convincing) for the claim that semantic features of some linguistic expressions and mental states depend on the physical or social environment. The arguments have to do with the properties of

[66] The literature on semantic externalism is vast. See Putnam (1975), Burge (1979b), and Davidson (1988) for arguments in favor of semantic externalism.

[67] For criticism of semantic externalism, see Boghossian (1989) and Segal (2000). For attempts to reconcile semantic externalism with knowledge of the contents of one's expressions and mental states, see Davidson (1987), Burge (1988), Kobes (1996), Gibbons (1996), Sawyer (1999), Heal (2001), Stueber (2002), and Hahwy (2002).

natural kind terms, our practices of attributing propositional attitudes, the individuation of concepts, and the way our expressions acquire meanings from the ways they are used. The contextual truth theorist has none of those justifications for her theory. The best she can do is say that it is one way of avoiding the liar's paradox, or even that it is better than its competitors. That does not count as a justification for the claim that the contents of some of our ordinary sentences that contain truth predicates are determined by empirical facts that seem to have nothing to do with the contents of our expressions.

That brings me to my second point. At least in the case of semantic externalism, the content-determining empirical facts are localized. Here are some examples. The contents of sentences of a language L that contain occurrences of 'water' depend on the chemical composition of the samples to which people who speak L have applied 'water'. The contents of sentences of a language M that contain occurrences of 'arthritis' depend on the norms present in the community of those who speak M that specify how 'arthritis' should be used. The contents of sentences of a language N that contain occurrences of 'tree' depend on the causal connections between trees and the speakers of N that use 'tree'. In the example from section three, all these factors are the same in w_1 and w_2. Thus, the contextualist cannot appeal to any of the factors that we commonly take to be relevant to the determination of content. Indeed, for any sentence at all, one can construct a sentence token that contains an occurrence of 'true' whose content depends on whether that sentence is true (according to the contextualist). Let ε be the sentence in question. Simply construct an example like the one in section two where the sentence token on inscribed on the blackboard in room 2 is 'either β is false or ε is true'. According to the contextualist, the content of α in a world where ε is true is different than the content of α in a world where ε is false. In the cases where semantic externalism applies, we need to know the chemical composition of 'water' or the norms of our linguistic community or our own causal history to be in a position to have semantic knowledge. In the case of contextualist theories of truth, there is no limit to what one might have to know in order to have semantic knowledge. The content of a sentence that contains a truth predicate might depend on *anything*.

The implications for our semantic knowledge from contextualist theories of truth are far more radical than those from semantic externalism. The aletheic contextualist has no justification for these radical consequences other than that they are the cost of solving the liar paradox. Finally, the arguments that are supposed to justify contextual theories of truth (e.g., the two-line puzzle and the strengthened liar reasoning) should be convincing only to contextualists. Therefore, there is no good reason to think that truth predicates are context-dependent and there are plenty of good reasons to think that they are not.

B.10 Conclusion

I have argued that the syntactic and semantic features of some sentence tokens do not determine whether they are paradoxical. I take it as a condition on theories of truth that they should be consistent with this empirical fact. Many are not. In particular I discussed five popular and influential doctrines that are incompatible with the riskiness of truth. I argued that the riskiness thesis is incompatible with: the view that the utterances of paradoxical sentences are not assertions, the claim that propositions are primary truth bearers, minimalist theories of truth-aptness, and exclusive T-schema deflationism. I also argued that the arguments for contextual theories of truth are circular, that these theories do not respect the Gricean condition on theories of context, and that they imply that we do not know the contents of many of the sentences that contain truth predicates.

APPENDIX C

Revision and Revenge

C.1 Introduction

The revision theory of truth is currently one of the three most prominent approaches to the liar paradox (the others being fixed-point theories and contextual theories).[1] Although the revision theory must be formulated in a language that is expressively richer than those to which it applies, its proponents claim that it does not fall prey to revenge paradoxes. I argue that it does face a revenge paradox, and this revenge paradox casts doubt on both its claim to be an acceptable approach to the liar paradox and its prospects for applying to natural languages.

C.2 The Revision Theory

On the revision theory, truth is a circular concept in the sense that the definition of truth is circular (i.e., the definiendum appears in the definiens). The revision theory of truth is grounded

[1] Herzberger and Gupta independently developed ancestors of the revision theory I describe; see Gupta (1982) and Herzberger (1982a, 1982b). The theory I describe is expounded in Gupta and Belnap (1993). In particular, I focus on their theory $T^{\#}$; see Gupta and Belnap (1993: 182-190, 210-229). See Kripke (1975) for an example of a fixed-point theory, and see Simmons (1993) for a contextual theory.

on a theory of circular concepts, which implies that the meaning of a word that expresses a circular concept is captured by a rule of revision. Given a hypothesis about such a word's semantic features, a rule of revision specifies a more accurate hypothesis about its semantic features. For example, the rule of revision for a circular predicate F implies that if we assume F has a certain extension, then we can determine a different extension for F. That is, the rule of revision specifies the extension of F under different assumptions about the extension of F. Although circular concepts do not have fixed semantic features, one can use the revision rule for a given circular concept to acquire information about its semantic features by considering its behavior during repeated applications. For example, if we begin with an arbitrary extension for F and apply the revision rule for F over and over, we generate a sequence of extensions for F. Roughly, if a certain object b always ends up in the extension of F and stays there through repeated applications of the revision rule to different starting extensions, then b satisfies F. Likewise, if a certain object c always ends up outside the extension of F and stays there through repeated applications of the revision rule to different starting extensions, then c does not satisfy F. We can say that b is *categorically* F and that c is *categorically* not F.

When used as a semantic theory for truth, the theory of circular concepts yields the revision theory for truth. On the revision theory, the T-sentences of a certain language L (i.e., sentences of the form: ⟨p⟩ is true-in-L iff p) are partial definitions of 'true-in-L'.[2] Given the set of partial definitions for 'true-in-L', the revision theory for truth determines a revision rule for 'true-in-L', which is used to specify the set of categorical sentences of L (those that are categorically true or categorically false) and the set of non-categorical sentences of L.

[2] Gupta and Belnap insist on distinguishing between T-sentences as definitional equivalences (where 'iff' is read as '$=_{df}$') and T-sentences as material equivalences (where 'iff' is read as '≡'). They accept all the T-sentences as definitional equivalences, but not all the T-sentences as material equivalences; see Gupta and Belnap (1993: 138-139).

Like the vast majority of approaches to the liar paradox, the language in which the revision theory for truth is formulated is expressively richer than the languages to which it applies.[3] In particular, it does not apply to certain languages that contain categoricalness predicates. One way to interpret this limitation is that the revision theory is restricted from applying to those languages that contain categoricalness predicates because it would imply that some of their sentences are both categorical and not categorical. When confronted with this objection, Gupta and Belnap write:

> It is a perennial problem with theories of truth that while they apparently resolve one paradox, they allow the generation of another, more vicious one. It may be objected that our account is not exempt from this. For example, a sentence such as (1) presents us with difficulty.
>
> (1) Either this sentence is not categorical or it is not true.
>
> If (1) is not categorical then it must be true (because its first disjunct is true) and hence categorical. On the other hand, if it is categorical then it is like the Liar (because its first disjunct is false) and hence noncategorical. In either case, we can deduce a contradiction. Our proposal, it may be said, can perhaps account for the *ordinary* versions of the Liar, but it cannot deal with the *strengthened* versions such as (1) above. The central problem raised by the paradoxes, it may appear, is left unresolved in our approach, (Gupta and Belnap 1993: 253).

I call this objection the *revenge objection* (what Gupta and Belnap call strengthened versions of the liar paradox are often called *revenge paradoxes* because with them the liar paradox takes its revenge on the theory in question by using a key concept of the theory to construct a new paradox).

Gupta and Belnap have two replies to the revenge objection. First, they argue that the argument given in the passage above (which I call *argument A*) is unsound because it relies on

[3] Gupta and Belnap admit that this is the case: "We have been concerned, for the most part, with languages whose only problematic element is the truth predicate. Even if these languages can be enriched, the strengthened versions of the paradoxes show that an adequate description of any of them requires, within our general framework, a richer metalanguage," (Gupta and Belnap 1993: 256).

the principle that all truths are categorical, which is not a consequence of the revision theory.[4] Their second reply is contained in the following passage:

> Another problem with the argument is that it assumes that the notion "categorical in L" is in L itself. This assumption can be accepted, but it cannot be assumed that "categorical in L" has an ordinary logic and semantics. On the contrary, we should observe that this notion is just as circular as truth: We can determine the categorical sentences of L *only on the basis of a prior hypothesis concerning the extension of* "*categorical in L*." … Because of this, the argument for the first horn does not establish that (1) is categorical. It shows only that if (1) falls outside the extension of "categorical" at one stage of the revision process, then it falls in the extension at the next stage. A similar claim holds for the second horn, (Gupta and Belnap 1993: 255-6).

They argue that the revenge objection assumes that the language in question (call it *L*) has a categoricalness-in-L predicate; if so, then categoricalness-in-L is a circular concept as well, and it does not obey the classical reasoning used in the objection.

They go on to argue that for a given language L, there is a hierarchy of categoricalness-in-L concepts, each of which is a circular concept:

> Just as the sentences 'the Liar is true' and 'the Liar is not true' are pathological in the one case, the sentences 'the Strengthened Liar is categorical' and 'the Strengthened Liar is not categorical' are pathological in the other. Hence, the concept of categoricalness that is appropriate for describing the behavior of the Ordinary Liar is not appropriate for the Strengthened Liar. To correctly describe the behavior of the latter we need to appeal to a higher-level notion of categoricalness. This higher-level notion would itself manifest paradoxical behavior in the presence of vicious reference. And we would account for it in the same way. The higher-level paradoxes would demand a still higher-level notion for their description, (Gupta and Belnap 1993: 256).

To sum up: Gupta and Belnap reply to the revenge objection by arguing that argument A is unsound both because it appeals to the claim that all truths are categorical and because it does not respect the fact that categoricalness is a circular concept. Moreover, they construct a

[4] Gupta and Belnap (1993: 255).

hierarchy of circular concepts of categoricalness with which one can classify the revenge paradoxes.[5]

C.3 Criticism

I argue that the Strengthened Liar (i.e., sentence (1) from the first quotation) does pose a problem for the revision theory and that Gupta and Belnap's two replies to it are inadequate. Fortunately, Gupta and Belnap have already done most of my work for me by proving the indefinability (in a language L) of categoricalness-in-L. In particular, they prove that an extension ($\mathcal{L}^+$) of the classical language of arithmetic that contains its own truth-in-$\mathcal{L}^+$ predicate (which is governed by the revision theory for truth) *cannot* contain a predicate that is true of all and only the categorical sentences of $\mathcal{L}^+$.[6] This theorem is analogous to Tarski's indefinability theorem for truth, which he proves with a *reductio* by deriving that a variant of the liar sentence is both true and not true.[7] Not surprisingly, Gupta and Belnap prove the indefinability of categoricalness-in-$\mathcal{L}^+$ by deriving that a variant of the Strengthened Liar is both categorical-in-$\mathcal{L}^+$ and not categorical-in-$\mathcal{L}^+$. They even comment: "The above proof is exactly parallel to the proof of Tarski's Indefinability Theorem. The principal difference between the two is that where the latter uses the Ordinary Liar, the former uses the Strengthened Liar," (Gupta and Belnap 1993: 230). To be clear: Gupta and Belnap *prove* that if a language L contains a categoricalness-in-L predicate, then the revision theory of truth implies that a certain sentence of L (e.g., a variant of the Strengthened Liar) is both categorical-in-L and not categorical-in-L (L must have other properties as well, but those

[5] See Gupta and Belnap (1993: 229-235) for the semantics of categoricalness.

[6] Gupta and Belnap (1993: 229-230).

[7] See Tarski (1933).

are irrelevant for my purposes). I call this subproof of their proof of the indefinability theorem *argument B.*[8]

Argument B shows that Gupta and Belnap's first reply to the revenge objection is inadequate. Argument A and argument B have the same conclusion: if the revision theory for truth is applied to a language L that contains its own categoricalness-in-L predicate, then it implies that a certain sentence of that language (e.g., a variant of the Strengthened Liar) is both categorical-in-L and not categorical-in-L. However, the arguments are quite different. In particular, argument A depends on the problematic principle that all truths are categorical, while argument B does not. Gupta and Belnap attribute argument A to the proponent of the revenge objection, and they claim that it is unsound.[9] I agree that it is unsound. However, there is a sound argument for the same conclusion; namely, their own argument B. I will not speculate on why they attribute an unsound argument to the objector when their own proof of the indefinability theorem contains a sound argument for the objector's claim.

What about their second response to the revenge objection? Gupta and Belnap assume that if a language L contains a categoricalness-in-L predicate, then this predicate expresses a circular concept. Given that Gupta and Belnap prove that categoricalness-in-L is *not* definable in L, it should seem odd that they claim that L *can* contain a categoricalness-in-L predicate. Moreover, they claim that if 'categorical-in-L' is in L, then it expresses a circular concept, but in

[8] Here is their proof of the indefinability theorem: "Suppose for *reductio* that $A(x)$ defines X in $\mathcal{L}^+$ [X is the set of Gödel numbers of sentences categorical in $\mathcal{L}^+$, which is the classical language of arithmetic extended with a truth predicate '$\mathbf{T}$', such that '$\mathbf{T}$' is governed by revision semantics]. Construct by the Gödelian techniques a 'fixed point' for the formula '$\sim A(x) \vee \mathbf{T}(x)$'; i.e., the sentence B such that (1) $B \leftrightarrow (\sim A(\ulcorner B \urcorner) \vee \sim \mathbf{T}(\ulcorner B \urcorner))$ is true in M + g, for all possible classical interpretation g of $\mathbf{T}$. Since $A(x)$ defines X, we know that (2) B is categorical in $\mathcal{L}^+$ if and only if $A(\ulcorner B \urcorner$ is valid in $\mathcal{L}^+$ and that (3) B is not categorical in $\mathcal{L}^+$ if and only if $\sim A\ (\ulcorner B \urcorner)$ is valid in $\mathcal{L}^+$. Now, either B is valid or $\sim B$ is valid or B is not categorical. Each of these assumptions can be shown to imply a contradiction. Suppose, for instance, that B is valid. Then $\mathbf{T}(\ulcorner B \urcorner)$ must also be valid. Since (1) is valid, and validity is closed under logical consequence, it follows that $\sim A(\ulcorner B \urcorner)$ is valid. This contradicts (2). The arguments for the other two cases are similar," (Gupta and Belnap 1993: 230; I have altered the use/mention conventions in this passage for economy).

[9] Gupta and Belnap attribute argument A to Cargile (1986) and Priest (1987); see Gupta and Belnap (1993: 253 n. 1)

their proof of the indefinability theorem (i.e., in argument B), they treat ‘categorical-in-L’ as if it obeys classical logic (i.e., as if it does not express a circular concept).

On reflection, we see that there is no tension between their claims about the definability and circularity of categoricalness and their proof of the indefinability theorem. The concept of categoricalness-in-L employed by a revision theory for a language L is different from the concept of categoricalness-in-L that can be defined in L by a revision theory. The concept of categoricalness-in-L employed by a revision theory is not definable in L, is not circular, and gives rise to revenge paradoxes, while the concept of categoricalness-in-L that can be defined in L by a revision theory for L is (obviously) definable in L, is circular, and does not give rise to revenge paradoxes. A sequence of examples should illustrate these facts.

Let L_0 be a first order language with a truth predicate ‘true-in-L_0’ and the usual capacity for self-reference (via arithmetization or names for its linguistic expressions). L_0 does not contain any categoricalness predicates. Let M_0 be a language in which a revision theory for truth-in-L_0 is formulated; call this theory T_0. Because T_0 employs ‘categorical-in-L_0’, T_0 is not expressible in L_0. Consider a different language, L_1, which is just like L_0, except that instead of containing ‘true-in-L_0’, it contains both ‘true-in-L_1’ and ‘categorical-in-L_1’. Let M_1 be the language in which a revision theory for truth-in-L_1 and categoricalness-in-L_1 is formulated; call this theory T_1.[10] T_1 employs a categoricalness-in-L_1 predicate; that is, it specifies which sentences of L_1 are categorical-in-L_1 and which ones are not categorical-in-L_1. It might seem that T_1 is expressible in L_1 or that the notion of categoricalness employed by T_1 is the same as the one expressible in L_1, but these impressions are inaccurate.

[10] One of the exciting features of Gupta and Belnap’s theory of circular concepts is that it can handle systems of interdependent concepts; T_1 is a revision theory for *both* truth-in-L_1 and categoricalness-in-L_1.

The concept of categoricalness-in-L_1 employed by T_1 is not circular, and it is a “genuine” categoricalness concept in the sense that it applies to all the categorical sentences of L_1 and fails to apply to all the non-categorical sentences of L_1. I use ‘categoricalness*-in-L_1’ as a term for this concept of categoricalness. Categoricalness*-in-L_1 is not definable in L_1, as the indefinability theorem shows. If L_1 contained a categoricalness*-in-L_1 predicate, then L_1 would contain a genuinely paradoxical variant of the Strengthened Liar (e.g., ‘either this sentence is false-in-L_1 or it is not categorical*-in-L_1’), as argument B shows (i.e., T_1 would imply that this sentence is both categorical*-in-L_1 and not categorical*-in-L_1).

On the other hand, the concept of categoricalness-in-L_1 that is expressible in L_1 is circular. I use ‘categoricalness$_1$-in-L_1’ as a term for this concept of categoricalness. Categoricalness$_1$-in-L_1 is not a “genuine” categoricalness concept because it is not the case that it both applies to all the categorical sentences of L_1 and fails to apply to all the non-categorical sentences of L_1 (e.g., it neither categorically applies nor fails to categorically apply to ‘either this sentence is false-in-L_1 or it is not categorical$_1$-in-L_1’).[11] Categoricalness$_1$-in-L_1 is, of course, definable in L_1. Consequently, L_1 contains a variant of the Strengthened Liar that expresses categoricalness$_1$-in-L_1 (i.e., ‘either this sentence is false-in-L_1 or it is not categorical$_1$-in-L_1’). However, this variant of the Strengthened Liar is not paradoxical (i.e., it is not the case that T_1 implies that it is both categorical*-in-L_1 and not categorical*-in-L_1). Instead, this sentence is not categorical*-in-L_1, but L_1 does not have the resources to express this fact.

Recall that Gupta and Belnap define a hierarchy of circular categoricalness concepts. To illustrate the relation between them and the notion of categoricalness employed by the revision theory, consider a language *L_2*, which is like L_1 and L_0 except that instead of the semantic predicates that belong to L_0 and L_1, L_2 contains a truth-in-L_2 predicate and two categoricalness

[11] That is, ‘either this sentence is false-in-L_1 or it is not categorical$_1$-in-L_1’ is not categorical*-in-L_1.

predicates: ‘categorical$_1$-in-L$_2$’ and ‘categorical$_2$-in-L$_2$’. Let T_2 be the revision theory for truth-in-L$_2$, categoricalness$_1$-in-L$_2$, and categoricalness$_2$-in-L$_2$. T$_2$ employs a concept of categoricalness-in-L$_2$ as well; call it *categoricalness*-in-L$_2$*. Categoricalness*-in-L$_2$ applies to all and only the categorical sentences of L$_2$ (i.e., it is a “genuine” categoricalness predicate); furthermore, it is indefinable in L$_2$ (as the indefinability theorem shows). Both categoricalness$_1$-in-L$_2$ and categoricalness$_2$-in-L$_2$ are circular concepts. They are obviously expressible in L$_2$, and the variants of the Strengthened Liar that express them (i.e., ‘either this sentence is false-in-L$_2$ or it is not categorical$_1$-in-L$_2$’ and ‘either this sentence is false-in-L$_2$ or it is not categorical$_2$-in-L$_2$’) are not paradoxical. The difference between categoricalness$_1$-in-L$_2$ and categoricalness$_2$-in-L$_2$ is that one can use categoricalness$_1$-in-L$_2$ to assert that the Liar is not categorical (by asserting that it is not categorical$_1$-in-L$_2$), and one can use categoricalness$_2$-in-L$_2$ to assert that ‘either this sentence is false-in-L$_2$ or it is not categorical$_1$-in-L$_2$’ is not categorical (by asserting that it is not categorical$_2$-in-L$_2$). That is a nice feature of L$_2$. However, L$_2$ does not have the resources to classify ‘either this sentence is false-in-L$_2$ or it is not categorical$_2$-in-L$_2$’.

With these facts about the categoricalness concepts in view, we can see that Gupta and Belnap’s second reply to the revenge objection is inadequate. They claim that categoricalness is a circular concept, that it can be described by a revision theory, and that it does not give rise to revenge paradoxes. Although they show how to construct circular concepts of categoricalness (e.g., categoricalness$_1$-in-L$_1$, categoricalness$_1$-in-L$_2$, and categoricalness$_2$-in-L$_2$), none of these concepts are genuine categoricalness concepts, none of them are employed by a revision theory, and none of them are expressed by the Strengthened Liar in the revenge objection. The revenge objection pertains to a language L that contains a categoricalness*-in-L predicate. Any such language contains a variant of the Strengthened Liar that expresses categoricalness*-in-L. If the revision theory applies to such a language, then it implies that the sentence ‘either this sentence is false-in-L or it is not categorical*-in-L’ is both categorical*-in-L and not categorical*-in-L. Thus, the worries Gupta and Belnap express in their formulation of the revenge objection hold true: (i) the revision theory *does* resolve one paradox while allowing for “the generation of another, more vicious one,” (ii) the revision theory “can perhaps account for the *ordinary* versions of the Liar, but it cannot deal with the *strengthened* versions,” and (iii) “the central problem raised by the paradoxes … is left untouched.”[12] The revenge objection stands.

C.4 Objections and Replies

Objection 1: The only notions of categoricalness are the circular ones that can be handled by the revision theory. Categoricalness*-in-L is incoherent.

Reply 1: The revision theory for language L employs categoricalness*-in-L. If this concept is incoherent, then the revision theory is incoherent as well. Recall the example of language L_1. T_1 is the revision theory for truth-in-L_1 and categoricalness$_1$-in-L_1. T_1 employs categoricalness*-in-L_1, which means that T_1 specifies which sentences of L_1 are categorical*-in-L_1 and which ones are not categorical*-in-L_1. T_1 does not employ categoricalness$_1$-in-L_1 (i.e., the circular concept of categoricalness that can be defined in L_1 by the revision theory) as Gupta and Belnap’s indefinability theorem shows.

Objection 2: Revision theories do employ circular concepts of categoricalness; the concept of categoricalness employed by a given revision theory T for a language L is one level

[12] All these quotes are from the first passage; Gupta and Belnap (1993: 253).

above the highest notion of categoricalness expressible in L. For example, L_1 contains a categroicalness$_1$-in-L_1 predicate; hence, T_1 employs categoricalness$_2$-in-L_1.

Reply 2: I have two replies to this objection. First, categoricalness*-in-L_1 (the concept employed by T_1) is not circular, while categoricalness$_2$-in-L_1 is circular. Thus, they are distinct concepts. That categoricalness$_2$-in-L_1 is circular follows from its definition. The concept of categoricalness*-in-L_1 is not circular because it is defined in terms of non-circular set-theoretic concepts.[13]

Second, argument B with 'categorical*-in-L_1' in place of 'categorical' is sound and shows that categoricalness*-in-L_1 is not definable in L_1 (so long as the assumptions of the indefinability theorem hold). However, argument B with 'categorical$_2$-in-L_1' in place of 'categorical' is invalid; hence, it does not show that categoricalness$_2$-in-L_1 is indefinable in L_1. Therefore, categoricalness*-in-L_1 is not identical to categoricalness$_2$-in-L_1. Similar reasoning shows that categoricalness*-in-L_1 is not identical to any of the circular categorical concepts. Again, similar reasoning shows that none of the revision theories employs a circular concept of categoricalness.

Objection 3: Either the notion of categoricalness for a language L is expressible in L or it is not. If it is expressible in L, then it is circular and the version of the Strengthened Liar that expresses it is not paradoxical. If the notion of categoricalness for L is not expressible in L, then it is not circular and the version of the Strengthened Liar that expresses it does not belong to L; hence, it is false (i.e., it attributes either falsity-in-L or non-categoricalness-in-L to itself, but it satisfies neither because it does not belong to L). The indefinability theorem shows that it is

[13] Categoricalness*-in-L_1 is defined in terms of validity on a set of definitions in a model for a theory of circular concepts, which is defined in terms of set theoretic constructions and truth-in-a-model (which is not circular); see Gupta and Belnap (1993: 145-147, 166-174, 182-189)

impossible for a language to contain a categoricalness predicate with which one can construct a version of the Strengthened Liar that is genuinely paradoxical.

Reply 3: On this objection, it is impossible for a language to contain its own categoricalness* predicate. This is obviously false. One can construct a language that contains its own categoricalness* predicate; however, the revision theory cannot apply to such a language while remaining consistent.[14] The indefinability theorem assumes the consistency of the revision theory. That is, given that the revision theory is consistent and that it applies to a language L, L cannot contain a categoricalness*-in-L predicate. Thus, given that the revision theory is consistent, and that it applies to L, L cannot contain a genuinely paradoxical version of the Strengthened Liar. The proponent of the revenge objection does not dispute these claims; however, she claims that either the revision theory is inconsistent or it applies only to languages that are expressively impoverished (e.g., those that do not contain their own categoricalness* predicates). Citing the indefinability theorem is not an adequate response to this objection. Indeed, the proponent of the revenge objection accepts the indefinability theorem and claims that its central argument (argument B) shows that the revision theory is either inconsistent or expressively restricted. In either case, the revision theory does not constitute an acceptable approach to the liar paradox.

C.5 Conclusion

The revision theory for truth faces a revenge paradox. Although Gupta and Belnap address this objection, their replies are inadequate. Their first reply is that the argument used in the objection

[14] For example, English (or my idiolect of English at noon GMT on January 1, 2004) contains its own categoricalness* predicate.

is unsound. They attribute an unsound argument (argument A) to the objector and point out that it is unsound; I agree with them. However, their own indefinability theorem is proved with the help of a sound argument for the objector's claim (argument B). Thus, their first reply is inadequate. Their second reply is that the revenge objection does not respect the fact that categoricalness is a circular concept; when one treats categoricalness as a circular concept, there is no revenge paradox. However, the concept of categoricalness employed by a revision theory (e.g., categoricalness*-in-L_1) is not circular, it gives rise to revenge paradoxes, and it is indefinable in the language in question. Although Gupta and Belnap show how to introduce categoricalness predicates (e.g., categoricalness$_1$-in-L_1) that do not give rise to revenge paradoxes, these are not "genuine" categoricalness predicates, and revision theories do not employ them.

The consequences of the revenge objection for the revision theory are severe. Gupta and Belnap "solve" the liar paradox only by using the familiar trick of restricting the expressive resources of the languages they consider. In particular, their theory is not applicable to a language (like English) that can express their theory.[15] At this point in our battles with the liar paradox, we know that solutions that incorporate this maneuver are unacceptable.

[15] I am not claiming that English is semantically self-sufficient (i.e., English can express the complete semantic theory for English) or that a theory of truth should be applicable to languages that are semantically self-sufficient. Given what I have said, Gupta and Belnap's attacks on semantic self-sufficiency criticisms are misplaced; see Gupta and Belnap (1993: 256-259). I claim that they cannot apply a revision theory to a language that contains its own categoricalness* predicate. English contains its own categoricalness* predicate. Hence, their theory does not apply to English.

APPENDIX D

PURPORTEDLY INTERNALIZABLE SEMANTIC THEORIES FOR TRUTH

D.1 INTRODUCTION

In Chapters One and Two, I introduced the concept of internalizability in an effort to categorize the relations between semantic theories, the languages to which they apply, and the languages in which they are formulated. In particular, I wanted to show that semantic theories for truth that do not apply to the languages in which they are formulated are unacceptable. In Chapter Two, I presented five internalizability requirements, and I argued that any acceptable semantic theory is internalizable for every language and that any acceptable semantic theory for truth is internalizable for every natural language to which it applies. A semantic theory that is internalizable for a single language does not require a substantive distinction between target language (i.e., the language to which it applies) and employed language (i.e., the language in which it is formulated). However, a semantic theory can be internalizable for one language without being internalizable for every language and without being internalizable for a single natural language. Thus, a semantic theory that does not require a substantive distinction between target language and employed language might not be internalizable for a natural language.

A number of semantic theories for truth have appeared whose proponents claim they do not require a substantive distinction between target language and employed language. These theorists claim that this feature qualifies these theories to apply to natural languages. In this paper, I compare and contrast these theories, and I evaluate them with respect to the various internalizability requirements. In section two, I define 'internalizability' and present three internalizability requirements. The semantic theories are divided into two groups. The first group of semantic theories for truth are based on consistency theories of truth—theories that imply that truth is a consistent concept. These include: Reinhardt's theory (which employs a formalist strategy), McGee's theory (which implies that truth is a vague concept), Simmons' theory (which implies that truth is context dependent), Field's theory (which implies that truth displays indeterminacy), and Maudlin's theory (which implies that truth is partially defined, and only grounded sentences have truth-values). I discuss each of these in section three. The second group of semantic theories for truth are based on inconsistency theories of truth—theories that imply that truth is an inconsistent concept. These include: dialetheism (which implies that some sentences are both true and false), Yablo's theory (which implies that truth is a particular type of circular concept), Elkund's theory (which implies that the constitutive principles for truth are inconsistent and that an acceptable assignment of semantic values should satisfy a weighted majority of these principles), and the semantic theory for truth I presented in Chapters Four, Five, Six, and Seven (which implies that truth is a confused concept with six components). I discuss each of these in section four.

D.2 Internalizability

For a full discussion of internalizability, the motivations for introducing this terminology, and my arguments for internalizability requirements on semantic theories, the reader should consult Chapters One and Two. In this section, I reproduce the definitions, and I formulate the requirements.

A semantic theory is *internalizable* for a given language if and only there exists an extension of the language such that the theory is expressible in the extension and the semantic theory applies to everything in the extension to which it is supposed to apply. The following is a more elaborate definition of internalizability:

> A semantic theory T that purports to specify the meanings of sentences that express a concept X is *internalizable for a language L* if and only if there exists an extension of L such that all the sentences that compose T can be translated into sentences that belong to the extension of L and T specifies the meanings of all the sentences of the extension of L that express X.

Below I provide definitions of 'semantic theory', 'language', 'expressible', 'applies', and related terms:

> A *theory* is a set of declarative sentences all of which belong to a single language. A *semantic theory* is a theory that specifies the meanings of certain sentences that belong to some particular language or languages. A specification of meaning by a semantic theory is an *assignment*.[1]
>
> A *language* is a function from sets of sentences (syntactic strings) to a set of sentential meanings. A language L_0 is a *sublanguage* of a language L_1 if and only if the set of sentences of L_0 is a subset of the set of sentences of L_1 and the set of sentential meanings of L_0 is a subset of the set of sentential meanings of L_1. A language L_1 is an *extension* of a language L_0 if and only if L_0 is a sublanguage of L_1.

[1] I use the locution 'semantic theory for X', where X is a placeholder for the name of a concept (e.g., a semantic theory for moral obligation, a semantic theory for truth). A semantic theory for X specifies the meanings of the sentences of certain languages that express the concept X (e.g., a semantic theory for truth specifies the meanings of sentences that express the concept of truth). In particular, a semantic theory for X that applies to a language L that can express X (an *X-language*) assigns meanings to the sentences of L that express X (the *X-sentences* of L). A *theory of X* is a theory that specifies the nature of X. A semantic theory for X is based on both a theory of meaning and a theory of X (see Chapter Three for discussion).

> A theory T that belongs to one language L_0 is *expressible* in another language L_1 if and only if for every sentence q that composes T, there exists a sentence p of L_1 such that p is a translation of q. A sentence of one language is a *translation* of a sentence that belongs to another if they have the same or relevantly similar meanings (or contents).
>
> A semantic theory T for X *applies to a language L* if and only if L is an X-language and T does not contain a restriction specifying that it does not provide the meanings for sentences of L. A semantic theory T for X *applies to a sentence S of L* if and only if S is an X-sentence and T does not contain a restriction specifying that it does not provide a meaning for S. A *restriction* for a semantic theory T is a claim that T does not provide meanings for the sentences of certain languages or that T does not provide the meanings of certain sentences of certain languages.

With these definitions in mind, we can define some concepts that are helpful in discussions of internalizability, and we can provide a more economical definition of internalizability (assume that T is a semantic theory for X and L is a language):

> T is *descriptively complete for L* if and only if T provides an assignment for every X-sentence of L.
>
> T is *descriptively complete* if and only if for every X-language L, T is descriptively complete for L.
>
> T is *descriptively correct for L* if and only if T is consistent and T provides a correct assignment for every member of L in its scope.
>
> T is *descriptively correct* if and only if for every X-language, T is descriptively correct for L.
>
> T is *internal for L* if and only if T is expressible in L and T is descriptively complete for L.
>
> T is *internalizable for L* if and only if there exists an extension L′ of L such that T is internal for L′.
>
> T is *essentially external for L* if and only if it is not the case that T is internalizable for L.

It turns out that a semantic theory for X can be internalizable for one language and essentially external for another. Furthermore, a semantic theory can be internalizable for a formal language (or a sublanguage of a natural language) without being internalizable for a natural language.

Thus, the definition of internalizability (which is a relation between a semantic theory and a language) permits the formulation of several different properties of semantic theories. I focus on three particular properties:

> (WEAK) A semantic theory T for X is *weakly internalizable* if and only if there exists an X-language L such that T is internalizable for L.
>
> (NATURAL) A semantic theory T for X is *naturally internalizable* if and only for every natural language L, T is internalizable for L.[2]
>
> (STRONG) A semantic theory T for X is *strongly internalizable* if and only if for every X-language L, T is internalizable for L.

With these definitions in mind, we can classify the various semantic theories for truth listed in the introduction.

Before discussing the individual theories, I want to discuss four topics: (i) the relation between internalizability, revenge paradoxes, and self-refutation problems, (ii) the relation between internalizability and language-specific truth predicates, (iii) using a natural language to describe that natural language, and (iv) semantic theories that imply that some linguistic expressions are unintelligible.

As I explain in Chapters One, Two, and Three, semantic theories for truth that are not internalizable for certain languages have been restricted from applying to certain languages or certain sentences. A semantic theory for truth that is not internalizable for some language has been restricted from applying to some sentences of that language in order to prevent either a revenge paradox or a self-refutation problem from rendering the theory false or unacceptable.

[2] I admit that the concept of natural internalizability is vague because the concept of a natural language is vague. However, I take English (or, my idiolect of English at noon GMT on January 1, 2005) as my paradigm of a natural language. An alternative definition of natural internalizability is ‘a semantic theory T for X is naturally internalizable if and only if there exists a natural language L such that T is internalizable for L’; however, if a natural language is just a language that is used by some community of people, then this notion of internalizability would not differ substantially from weak internalizability, because most any language can be used by some community of people.

(See Chapters One, Two, and Three for discussion of revenge paradoxes and self-refutation problems.)

The second topic I want to discuss is language-specific concepts of truth. Most semantic theories for truth are actually semantic theories for language-specific concepts of truth. A language-specific concept of truth is like the concept of truth except that it is satisfied only by sentences of a single language. For example, ‘true-in-English’ is a language-specific truth predicate. Any sentence that is true-in-English is true, but true sentences of other languages are not true-in-English. One reason for focusing exclusively on language-specific (LS) truth predicates is that it is easier to avoid the liar paradox and its relatives for an LS truth predicate. By considering only LS truth predicates, one can ignore liar-like paradoxes that result from inter-linguistic truth attributions (e.g., if p is the German sentence ‘q ist wahr’—which means that q is true—and q is the English sentence ‘p is false’, then both p and q are paradoxical). In addition, it is easier for an account of an LS concept of truth to avoid revenge paradoxes. If T is a theory of truth and T implies that paradoxical sentences are truth-value gaps, then a revenge paradox for T concerns the sentence (λ′), which is ‘(λ′) is either false or a gap’; T implies that (λ′) is both true and either false or a gap. However, if T′ is a theory of truth-in-L (where ‘L’ is the name of a language), and T′ implies that paradoxical sentences of L are truth-in-L-value gaps, then T′ does not face a revenge paradox so long as L does not contain a gap predicate. That is, if L does not contain a gap predicate, then (λ″), which is ‘(λ″) is either false-in-L or a gap-in-L’ is not a member of L. Hence, (λ″) is not true-in-L. Thus, (λ′) does not constitute a revenge paradox for T. Therefore, a theorist can avoid revenge paradoxes by constructing a theory of an LS truth predicate and insuring that the language in question is expressively impoverished.

In Appendix A, I argue that there is no good way to explain natural language truth predicates in terms of LS truth predicates. This result affects internalizability in the following way. Assume that T is a semantic theory for an LS concept of truth (truth-in-L). Assume that T is descriptively correct for L, that T is descriptively complete for L, and that T is expressible in L. Hence, T is internalizable for L. A theorist who accepts T might claim that T is naturally internalizable as well. Of course, T is a semantic theory for an LS concept of truth, and the most common concept of truth expressible in a natural language is not an LS concept of truth. However, the theorist who accepts T could claim that natural language concepts of truth can be explained in terms of LS concepts of truth. If that is correct, then T might turn out to be naturally internalizable (e.g., if it is both descriptively complete for the natural language and expressible in the natural language). However, my claim that natural language truth concepts cannot be explained in terms of LS concepts of truth blocks this move. Because T is a semantic theory for LS truth concepts, either T does not apply to natural languages or it is not descriptively correct for natural languages. Either way, T is not both naturally internalizable and descriptively correct.

The third topic I want to discuss is natural languages. The conventional wisdom is that a necessary condition on a semantic theory for truth that does a good job of describing a natural language is that the theory does not require a substantive distinction between employed language and target language. The idea is that if we are using our natural language to describe our natural language, then a theory that requires a substantive distinction of this sort is unacceptable. This view has led several philosophers to construct a semantic theory for truth and a formal language such that the semantic theory for truth is internal for that formal language. The common assumption is that because these theories apply to a language in which they are formulated, these theories will do a good job of describing a natural language. Although I agree that a semantic theory for truth that requires a substantive distinction between employed language and target language is unacceptable for use on a natural language, the problem with the above reasoning is that although these theorists propose semantic theories for truth that are internalizable for the formal language in which they are formulated (i.e., they are weakly internalizable), such theories are often not internalizable for a natural language (i.e., they are not naturally internalizable). If we are to use our natural language to describe our natural language, then we need a naturally internalizable semantic theory, not one that is merely weakly internalizable. I argue that this is a problem for many of the theories that purport to be internalizable: although they might be weakly internalizable, they are not naturally internalizable, and weak internalizability is not enough.[3]

I rely on two claims in my evaluation of semantic theories for truth. First, any acceptable semantic theory for truth is naturally internalizable.[4] Second, any acceptable semantic theory for truth is descriptively correct for every language to which it applies. Several theorists attempt to insure that their semantic theories of truth are internalizable for natural languages by claiming that certain linguistic expressions (e.g., non-monotonic sentential operators) are unintelligible; they then argue that their theories need not apply to languages with such expressions or to sentences that contain such expressions. I call this move an *unintelligibility maneuver*. Although it is unclear what these theorists mean by 'unintelligible', I take it that they mean either that these linguistic expressions are meaningless or that they express inconsistent concepts.

[3] It seems to me that a semantic theory might be naturally internalizable without being strongly internalizable; however, it also seems to me that it would be difficult to argue that a semantic theory has this status because any such argument would have to specify the relevant linguistic resources that are present in some languages (but not natural languages). Any such specification would take place in a natural language, which would render the claim false.

[4] See Chapter Two for an argument for this condition.

I take it for granted that if there is an established practice of using a linguistic expression, then that linguistic expression is meaningful.[5] For each of the linguistic expressions that are labeled unintelligible by these theorists, there is an established practice of using them. Moreover, these linguistic expressions belong to some natural languages. Thus, any semantic theory for truth that incorporates an unintelligibility maneuver is unacceptable. To see why, assume that L is a natural language with a non-monotonic sentential operator and that T is a semantic theory for truth that applies to L and implies that non-monotonic sentential operators are unintelligible. If the theory implies that such expressions are meaningless, then it is not descriptively correct for L, because its assignments to the truth-sentences of L that contain this expression are incorrect— the sentences of L that contain this expression are meaningful, but the theory implies that they are meaningless. If the theory implies that such expressions express inconsistent concepts, then it is not descriptively complete for L, because the truth-sentences of L that contain these expressions are outside its scope. Because any extension of L will contain these expressions as well, for every extension L′ of L, the theory is not descriptively complete for L′. Hence, the theory is not internalizable for L, because if a theory is internalizable for a language, then it is descriptively complete for some extension of that language. Therefore, if a semantic theory for truth incorporates an unintelligibility maneuver, then it is not both descriptively correct and naturally internalizable.

[5] Even linguistic expressions like ‘tonk’ are meaningful; they just express inconsistent concepts; see Prior (1960) for a discussion of ‘tonk’.

D.3 Consistency Theories

All the semantic theories in this section are based on theories of truth that imply that truth is a consistent concept. I discuss the distinction between consistent and inconsistent concepts in Chapter Four. In Chapter Three, I argued that any acceptable theory of truth that respects the truth rules and implies that truth is a consistent concept is restricted in such a way that any semantic theory for truth that is based on this theory of truth is not naturally internalizable. Dividing the purportedly internalizable semantic theories for truth into consistency theories and inconsistency theories should help the reader verify this conclusion.

D.3.1 Formalism (Reinhardt)

The first theory I consider is Reinhardt's, which he published in 1986.[6] Reinhardt is committed to providing a theory of truth that applies to the language in which it is formulated.

> Let us suppose, as I believe is intuitively correct, that one of the primary features of [truth] is that it is one notion: in particular it does not split into some hierarchy of notions. … Let us explain that the truth predicate of our formal language (call the language L) is intended to be taken in the sense of our preexisting informal notion of truth. … Unless we are prepared to entertain splitting the notion of truth, we are forced to admit that the metalanguage is included in the object language. If the formal language is to provide an adequate explication of the informal language that we use, it must contain its own metalanguage. I take it that this is in fact a desideratum for success in formulating a theory of truth, (Reinhardt 1986: 227-228).

Recall that I used this quote in Chapter One to motivate my account of internalizability.

The theory Reinhardt proposes is an extension and a reinterpretation of Kripke's theory of truth (for discussion of Kripke's theory, see Chapter One). In particular, Reinhardt proposes a different way of interpreting partial truth predicates. He argues that, although Kripke intends to

[6] Reinhardt (1986).

interpret truth as a partial predicate, the fixed-point model treats it as a completely defined predicate. Reinhardt links this problem to the fact that Kripke's theory of truth must be formulated in a language that is expressively richer than the ones to which it applies if it is to be consistent.

Reinhardt's approach uses a significance predicate, which he takes to be synonymous with 'is true or false'. He does not assume that insignificant sentences are meaningless, but he does propose a particular reading of them according to which they are to be treated as merely formal (in Hilbert's sense). They are to be treated as strings of symbols that can be manipulated, but that have no meaning outside this practice. Reinhardt requires that the fundamental principles of his theory turn out to be significant. Moreover he argues that this condition is equivalent to the condition that the theory of truth should apply to the language in which it is formulated.

Reinhardt uses the axiomatic theory of truth, KF, which is Feferman's axiomatization of Kripke's model theoretic semantic theory.[7] From KF, one can prove the schema $\langle S(\langle p\rangle) \rightarrow (T(\langle p\rangle) \leftrightarrow p)\rangle$, where $\langle S(\langle p\rangle)\rangle$ says that sentence $\langle p\rangle$ is significant and $\langle T(\langle p\rangle)\rangle$ says that $\langle p\rangle$ is true. Reinhardt uses KF to state two sufficient conditions on his theory of truth. Specifically, if KF $\vdash T(\langle p\rangle)$[8], then $\langle p\rangle$ is true, and if KF $\vdash S(\langle p\rangle)$, then $\langle p\rangle$ is significant. Thus, Reinhardt uses KF's pronouncements on truth and significance as a basis for his accounts of truth and significance. Note that he denies that if KF $\vdash \langle p\rangle$, then $\langle p\rangle$ is true; he takes as true the sentences for which KF proves truth ascriptions, not those that KF proves. Reinhardt completes his theory by adding the claim that if KF $\vdash T(\langle p\rangle)$ then $\langle p\rangle$ is true, to KF itself. That is, KF$^+$ is KF with the additional axiom $\langle T(\langle \forall x(\Theta(x) \rightarrow T(x))\rangle)\rangle$, where '$\Theta(x)$' is an arithmetical predicate.

[7] Feferman (1982).

[8] The turnstile is a symbol for the proof theoretic consequence relation.

The result is that, although Reinhardt's theory is consistent and proves a number of important facts about truth, it also proves results that, according to it, are not significant. Indeed, the axioms of KF are not significant. In particular, any claim that some significance attribution is insignificant turns out to be insignificant. Reinhardt's solution to this problem is to claim that these are uninterpreted symbols that have a use in deriving the results of the theory, but have no significance in themselves. Moreover, Reinhardt takes great care in the exposition of his theory to avoid asserting any insignificant sentences. The result is a clever presentation of an interesting thesis.

Is Reinhardt's theory internalizable in any of the three senses? Assume that Reinhardt's theory is the set of axioms of KF^+ and L is the formal language in which Reinhardt formulates it. KF^+ is expressible in L and, given that KF^+ is formulated in an artificial language whose sentences are stipulated to have the semantic properties they have and that the theory implies that they turn out to have these semantic properties, KF^+ is descriptively correct for L; hence, KF^+ is weakly internalizable.

Reinhardt achieves weak internalizability for his theory by claiming that many seemingly significant sentences are insignificant. One should not assume that this is an unintelligibility maneuver since 'significant' for Reinhardt is synonymous with 'neither true nor false'. His theory does face both a revenge paradox and a self-refutation problem though. These insure that it is not both descriptively correct for a natural language and naturally internalizable. The self-refutation problem comes from the fact that the theory has consequences that the theory implies are insignificant. Reinhardt readily admits this fact, but he does not discuss the self-refutation problem.[9] The revenge paradox concerns the fact that any attempt to add the resources that

[9] Reinhardt (1986: 236-7).

allow one to make a significant comment on the insignificance of a sentence renders the theory inconsistent. Consider the following sentence:

(1) (1) is either false or insignificant.

Reinhardt's theory implies that (1) is both true and either false or insignificant. To avoid this result, he restricts his theory so that it does not apply to languages with insignificance predicates. Reinhardt comments, "[t]he most awkward point in carrying this out is translating our use of 'non-significant'. This does not usually mean 'is not significant'.... It generally means 'is not in X', where X is some reasonably comprehensive collection of significant sentences. I may perhaps be justly accused of replacing a hierarchy of truth predicates with a hierarchy of non-significance predicates," (Reinhardt 1986: 228). Given that natural languages contain insignificance predicates, Reinhardt's theory is not both descriptively correct and naturally internalizable. Moreover, given that the theory is self-refuting (i.e., it implies that some of its consequences are insignificant), if one has access to a language with more expressive resources, one can argue that this fact renders it false (see Chapter Three for a discussion of self-refutation problems). Nevertheless, Reinhardt's theory is, as far as I know, the first theory of truth to permit a descriptively correct, weakly internalizable semantic theory for truth. That alone is a significant accomplishment.

D.3.2 VAGUENESS (MCGEE)

The second theory I consider was published by McGee in 1991 (although an abstract of the theory was published in 1989).[10] McGee, like Reinhardt, assumes that a theory of truth that

[10] McGee (1989, 1991).

requires a substantive distinction between employed language and target language is unacceptable.

> It must be possible to give the semantics of our language within the language itself. … [This requirement] is intended to hold open the possibility that the methods we develop can be applied to natural languages. If in developing the theory of truth for a language, we required the services of an essentially richer metalanguage, that possibility would be closed off. … [It] makes it reasonable to hope that our methods can be used to get a semantics of a natural language, (McGee 1991: 159).

Recall that I used this quotation in Chapter One to motivate my account of internalizability. McGee claims that our naïve conception of truth is inconsistent, where the naïve conception of truth is governed by the truth rules. Although McGee offers no account of inconsistent concepts, he assumes that they should be replaced and he sets out to provide a replacement.

McGee's theory of truth is subtle and ingenious, and I do not have the space to describe it in detail. Because my only concern in this paper is internalizability, I give a rough sketch of the theory below that is sufficient for my purposes. McGee's replacement concept of truth is vague—the sentences that are overdetermined on the naïve conception of truth are underdetermined on McGee's conception. In order to arrive at his theory of truth, McGee uses a supervaluation–based theory of vague concepts and applies it to truth. Like many accounts of vagueness, McGee introduces a 'definite' operator to distinguish unproblematic from problematic cases of application. For example, one might say that someone is definitely bald, in which case, the person does not fall within the borderline between baldness and nonbaldness. Likewise, McGee distinguishes between truth and definite truth. On McGee's theory, the ascending and descending truth rules preserve definite truth. That is, the following principles hold:

If ⟨p⟩ is definitely true, then ⟨⟨p⟩ is true⟩ is definitely true.

If ⟨p⟩ is definitely not true, then ⟨⟨p⟩ is true⟩ is definitely not true.

If ⟨p⟩ is unsettled, then ⟨⟨p⟩ is true⟩ is unsettled.

For his theory of definite truth, McGee uses Kripke's fixed-point theory. By appealing to the notion of a partially interpreted language, McGee proves that both his supervaluation semantic theory for truth and his fixed point semantic theory for definite truth apply to the language in which they are formulated. One of the keys to this result is that the formal language in which his theories are formulated does not contain sentences that pose revenge paradoxes. That is, the sentence:

(2) (2) is false or unsettled,

is unsettled, and so not definitely true, but it is not *definitely* unsettled, thus, no paradox results from it.[11] Furthermore, because (2) is not a consequence of either theory, it does not pose a self-refutation problem.

Is McGee's theory internalizable in any of the three senses? One must be careful in answering this question to distinguish between his theory of truth and his theory of definite truth. One thing is for sure, McGee proves that both theories are internalizable for the formal language in which they are formulated and that they are descriptively correct for this language. Thus, a version of each theory is weakly internalizable and descriptively correct. He achieves this result by setting up his theories so that 'if p is definitely true, then p is true' is not definitely true. This is a counterintuitive result, but he needs it to ensure that (2) does not pose a revenge paradox. He also has to deny that the vague concept of truth he presents can be made more precise. Attempts at precisification result in revenge paradoxes.[12] In addition, he restricts the ascending and

[11] McGee (1991, ch. 9).

[12] See Yablo (1989), Priest (1992), Simmons (1993), Priest (1994), Tappenden (1994), and Mills (1995) for discussion of this aspect of McGee's theory.

descending truth rules to categorical contexts—they are invalid in hypothetical contexts like conditional proofs. Thus, McGee's theory does not respect the truth rules.

Is McGee's theory of truth or his theory of definite truth naturally internalizable? It seems to me that neither one is for the following reasons. The theory of definite truth employs a fixed-point semantic theory that is similar to the one Kripke proposes. Fixed-point semantic theories are restricted to languages with monotonic sentential operators.[13] Thus, his theory of definite truth is restricted so that it does not apply to languages, like English, that contain non-monotonic sentential operators (e.g., exclusion negation). For such languages, the construction never reaches a fixed point, and so it does not have any assignments at all. Thus, McGee's theory of definite truth is not naturally internalizable. His theory of truth is not naturally internalizable either. The notion of definite truth employed by the theory of truth is vague (so that it avoids revenge paradoxes). However, in a language with a completely defined, non-partial definite truth predicate (which one can define with the help of a non-monotonic sentential operator) one can formulate a revenge paradox for McGee's theory. Thus, his theory must be restricted so that it does not apply to such languages. Nevertheless, McGee's theory of truth is the first to be weakly internalizable, descriptively correct, and not self-refuting.[14]

[13] Gupta and Martin (1984) prove that a fixed point semantics can handle certain non-monotonic linguistic devices (e.g., gap predicates), but they do not show that one can handle exclusion negation or other non-monotonic sentential operators.

[14] McGee's theory is a theory of LS truth concepts. He does not discuss the relation between these concepts and natural language truth concepts explicitly with respect to his theory of truth, but he does claim in other work that natural language truth predicates can be explained in terms of LS truth predicates; see McGee (1993). If we add this claim to his theory of truth it is not both descriptively correct and naturally internalizable.

D.3.3 Context-Dependence (Simmons)

Simmons published a theory of truth in 1993, called the *singularity theory*, that is quite different from those of Reinhardt and McGee.[15] Simmons's theory is in the tradition of contextual theories of truth—it implies that 'true' is a context dependent expression in the sense that its extension and anti-extension differ from context to context. Like Reinhardt and McGee, he claims that it does not require a substantive distinction between employed language and target language.

> I argue that the singularity proposal satisfies the criteria of adequacy developed earlier. In particular, I argue that my proposal does justice to Tarski's intuition that natural languages are semantically universal, in a way that is not undermined by diagonal arguments. I show that the singularity theory can accommodate not only our ordinary uses of 'true', but also the very semantic notions in which the theory is couched: the notions of *groundedness*, *truth in a context*, and *singularity*. Moreover, on the singularity account, the language of the theory does not stand to the object language as a Tarskian metalanguage to object language. Indeed, an ordinary use of 'true' in the object language includes in its extension the sentences of the theory, since these theoretical sentences are not identified as singularities. I conclude that the singularity theory respects the intuition that a natural language like English is semantically universal, (Simmons 1993: xi).

For Simmons, a language is *semantically universal* if and only if every semantic concept is expressible in the language. He claims that many natural languages are semantically universal and that an adequate theory of truth should apply to such languages. He also criticizes most other theories of truth for failing to meet this condition. (I drew the distinction between weak internalizability and natural internalizability in section one; one can find a similar distinction at work in Simmons' criticisms.[16])

[15] Simmons (1993, 1994, 2000, 2003).

[16] See Simmons (1993: chs. 3 and 4).

Simmons's theory is different from most contextual theories in that he does not use a hierarchy of contexts.[17] Each contextual use of 'true' is at the same level as every other one. However, each contextual use of 'true' has some gaps—sentences that are neither true nor false in that context; he calls these *singularities*. He provides a set of rules for determining which singularities are appropriate for which contexts and he is explicit about the fact that each contextual use of 'true' should be thought of as applying to all truth apt truth bearers except the singularities.[18]

Simmons provides a way to introduce 'groundedness', 'singularity', and 'context' predicates into the object language (i.e., the language containing the context dependent concept of truth). This is quite a feat given the difficulty that other approaches have had with these issues. Of course, they give rise to revenge paradoxes (e.g., a sentence that says of itself that it is not true in any context). Simmons then moves to a hierarchy of metalanguages to handle these cases. Each metalanguage contains a *context-independent* truth predicate that applies to the object language and all the metalanguages "below" it. An interesting aspect of his account is that all the sentences of the metalanguages are within the scope of the context dependent truth predicate, which is in the object language. Thus, in any given context, the sentences of the metalanguages are in either the extension, the anti-extension, or the gap-set of the context dependent truth predicate, which belongs to the object language. Furthermore, Simmons claims that the sentences that compose the singularity theory itself are within the scope of the context dependent truth predicate. Moreover, these sentences are not singularities in any context. Thus, in one sense, the theory describes a truth predicate that can apply to the sentences of the theory.

[17] See Parsons (1974), Burge (1979a), Barwise and Etchemendy (1987), Koons (1992), Gaifman (1992, 2000), and Glanzberg (2003) for other examples of contextual theories of truth. I discuss such theories in Appendix B.

[18] Simmons claims that his theory is not restricted to LS truth predicates.

Presumably, this fact is what leads Simmons to claim (in the passage quoted above), that the singularity theory can accommodate the notions it employs.

Is Simmons' theory of truth internalizable in any of the three senses? It seems to me that he answer is "no." Simmons' theory depends on a contextual account of the truth predicate to solve the liar. His solution depends on a hierarchy of metalanguages to solve the revenge liars that result from allowing the language to express the concepts of groundedness, singularity, and context dependence (concepts involved in the contextual account of truth). Of course, the theory also employs each of the context independent truth predicates that occur in the hierarchy of metalanguages. The question arises: in which language can the singularity theory be formulated? It seems that it cannot be formulated in any of the metalanguages because these lack the resources for constructing the hierarchy itself. Thus, if it can be expressed at all, then it must be expressible in the object language (the one with the contextual truth predicate). Can it be expressed here? I have my doubts. The object language would have to have enough resources to construct the hierarchy of metalanguages; in particular, it would have all the context-independent truth predicates that occur in the metalanguages. It seems to me that if it has these resources, then there is no sense in which they constitute a hierarchy of metalanguages. It seems instead that there is a hierarchy of context invariant truth predicates that belong to the object language. Then the question (or at least one question) would be: does this construction still avoid the revenge liars? It does not. Indeed, Simmons' combination of contextual truth predicate in the object language with context-invariant truth predicates in the hierarchy of metalanguages is designed to avoid revenge paradoxes because there is no "highest metalanguage" out of which one can diagonalize. However, the singularity theory itself cannot be expressed in any of these languages. If, instead, Simmons uses an object language with a single contextual truth predicate and an infinite hierarchy of context-invariant truth predicates, then there will be revenge liars in the offing. Thus, either his theory is inexpressible in the languages it posits or it is open to revenge paradoxes, which require it to be restricted to keep it consistent. Thus, it is not even weakly internalizable.

D.3.4 INDETERMINACY (FIELD)

In a series of recent papers, Field introduces an impressive account of partially defined expressions. From this account he derives a theory of truth and indeterminacy, a theory of vagueness, and a theory of properties.[19] He uses his account of partially defined expressions to provide a novel, powerful, and unified solution to the liar paradox, Curry's paradox, the sorites paradox, and the property version of Russell's paradox.[20] To accompany his account, he presents a new formulation of deflationism, a new non-classical logic with an intuitive conditional, and a non-standard probability calculus that allows him to explain degrees of belief in propositions that display indeterminacy.[21] It should be clear that a discussion of this entire account is beyond the scope of this paper. Instead, I focus on his theory of truth and

[19] See Field (2002, 2003a, 2003b, 2003c, forthcoming a, forthcoming b) for the theory of truth and indeterminacy, Field (2003b, 2003c) for the theory of vagueness, and Field (2003c, 2004) for the theory of properties.

[20] Curry's paradox is that, from intuitive assumptions, one can use the sentence 'if this sentence is true, then 0 = 1' to derive that 0 = 1 (or any other absurdity). The sorites paradox is that, from intuitive assumptions, one can derive that a person with a full head of hair is bald (the reasoning works for most vague expressions, not just 'bald'). The property version of Russell's paradox is that, from intuitive assumptions, one can derive that the property of non-self-instantiation both instantiates itself and does not instantiate itself. It seems that Field's account could be adapted to provide solutions for the paradoxes of denotation (Richard's paradox, Berry's paradox, and König's paradox) and Grelling's paradox of predication as well, but he does not emphasize this aspect of his work. Field is pessimistic about using it as a solution to the set-theoretic paradoxes; however, he also does not think that a new solution is in order. See Field (2004).

[21] See the previously cited works on his theory of truth for the first two topics and Field (2000, 2001b, 2003b, forthcoming c) for the non-standard probability calculus. One should be aware that Field presents two nonstandard probability calculi, one that is classical and the other nonclassical. He now endorses only the nonclassical version; see Field (2003c: 462).

indeterminacy, but I provide only the details that are relevant to deciding whether it is internalizable.

Field advertises his theory of truth as a "revenge-immune solution to the semantic paradoxes," and he claims that his semantic theory applies to the very language in which it is formulated.[22] Thus, he certainly claims at least weak internalizability for his theory. Before evaluating this claim, I want to describe his theory.

Field begins with a version of Kripke's theory of truth. A problem with Kripke's approach is that the languages he considers lack an intuitive conditional. One of Field's innovations is showing how to add a conditional to these languages that obeys most of the principles we associate with conditionals (e.g., it obeys modus ponens, and 'A → A' is a logical truth). Field's method for adding a truth predicate and a conditional to a first-order language results in a language that has truth-value gaps and for which the principle of excluded middle (i.e., $\ulcorner p \vee \sim p \urcorner$) fails in general.[23] The conditional behaves just like a material conditional if one assumes the relevant instances of the principle of excluded middle. According to Field, truth displays indeterminacy in the sense that some sentences (e.g., the liar sentence) are in neither the extension nor the anti-extension of 'true'; however, neither 'the liar sentence is not true', nor 'the liar sentence is not false' is a member of the extension of 'true' because the sentence 'the liar sentence is either true or false' is indeterminate. Instead, one can describe the status of the liar sentence by asserting that it is not determinately true and not determinately false (i.e., the liar sentence is indeterminate). One of the most satisfying aspects of Field's theory is that he

[22] Field (2003a, 2003b).

[23] Field denies that the languages he considers have truth-value gaps; see Field (2003b: 270). However, Field means something different by 'true-value gap' than most of those who work on truth and the liar paradox. The common usage is that a sentence p of a language L is a truth-value gap if p is a member of neither the extension nor the anti-extension of 'true-in-L'. On Field's usage, a sentence p of L is a truth-value gap if the sentence 'p is not true-in-L and p is not false-in-L' is in the extension of 'true-in-L'. The languages Field considers allow truth-value gaps in the common sense of the term, but do not allow truth-value gaps in the Fieldian sense of the term.

provides the means for asserting that the liar sentence is indeterminate in the languages he considers. He defines a sentential determinacy operator, ‘D’, in terms of his conditional:

$DA =_{df} A \wedge (\top \rightarrow A)$.

Here, ‘$\top$’ is any tautology (e.g., ‘$A \rightarrow A$’ or ‘$0 = 0$’).

That Field provides a characterization of the liar in the languages he considers is a huge advance over Kripke’s theory, which implies that the liar is a truth-value gap, but cannot be applied to languages that contain truth-value gap predicates. The problem with applying Kripke’s theory to such languages is that one can use a truth-value gap predicate to construct a revenge paradox for his theory. Given that Field’s theory implies that the liar sentence is indeterminate (i.e., not determinately true and not determinately false), the revenge paradox for Field’s theory of truth involves the sentence:

(3) (3) is either false or indeterminate.

Field is ready for it. Although the sentence ‘(3) is determinately true or determinately false’ is indeterminate, Field’s determinacy operator iterates non-trivially so that one can say that (3) is not *determinately* determinately true and not determinately false (determinate determinate falsity is equivalent to determinate falsity, but determinate determinate truth is stronger than determinate truth).[24] Of course, one can formulate a determinate determinate liar, but Field’s determinacy operator iterates again so that he can characterize it in the language as well. In fact, Field shows how to define a transfinite hierarchy of determinacy operators (D^{σ}) in terms of which he defines a transfinite hierarchy of determinate truth predicates (D^{σ}true(x)) and a transfinite hierarchy of indeterminacy predicates ($\sim D^{\sigma}$true(x) $\wedge$ $\sim$Dfalse(x)). (In other words,

[24] It might be helpful to see that ‘determinately determinately A’ is synonymous with ‘$(A \wedge (0 = 0 \rightarrow A)) \wedge (0 = 0 \rightarrow (A \wedge (0 = 0 \rightarrow A)))$’.

Field's theory incorporates what I called in Chapter Three the *robust response* to the revenge problem.)

When I say that Field constructs a transfinite hierarchy of determinacy operators, I mean that, given a system of ordinal notations (roughly, a certain mapping from a set of integers onto a segment of the ordinals), Field shows how to construct an operator $\ulcorner D^{\sigma} \urcorner$ for any ordinal σ in a proper initial segment of the recursive ordinals. There is no maximal recursively related system of ordinal notations; hence, given a system of ordinal notations for a proper initial segment of the recursive ordinals, one can construct a new system of ordinal notations for a larger proper initial segment of the recursive ordinals. One can construct a new set of determinacy operators based on the new system of ordinal notations such that there exists an operator $\ulcorner D^{\sigma} \urcorner$ in the new set that is not a member of the old set. In less precise language: there are many different ways of constructing the hierarchy of determinacy operators, and there is no "highest" hierarchy of them—given one hierarchy of determinacy operators, one can construct a "higher" one.[25]

Is Field's theory internalizable in any of the three senses? I grant that Field's theory can be formulated without employing notions outside his target language; thus, it is weakly internalizable. He argues convincingly that he appeals only to a model-relative concept of semantic value to set up his theory and to prove that his theory is consistent in the non-classical logic he provides.[26] Moreover, it seems to me that none of the consequences of his theory that are expressible in the target language are counted as untrue by the theory. Thus, his theory does not face any self-refutation problems. Is his theory naturally internalizable? The answer is "no." The most obvious reason is that Field's theory depends on a fixed-point construction that is

[25] See Field (2003a, 2003b, forthcoming b: fn. 14, fn. 21). I want to thank Hartry Field for conversations on this aspect of his theory in which he pointed out several mistakes in my exposition. See Rogers (1967: 205-213), and Setzer (1999), and Rathjen (1999) on ordinal notations.

[26] Field (2003a: 167-176; 2003b: 302-305).

similar to Kripke's theory. If he begins with a target language that contains an exclusion negation operator, then the construction will never reach a fixed point. Thus, his theory does not apply to languages, like English, that can express exclusion negation.

Field's response to this objection would almost certainly be to claim that a sentential operator like exclusion negation is unintelligible.[27] If we grant him this claim (and extend it to any linguistic device that prevents the construction from reaching a fixed point), then his theory is naturally internalizable, and indeed, it is strongly internalizable. However, I have argued that we should not accept his unintelligibility maneuver (see section two of this paper and Chapter Three).

A related point is that one can construct a revenge paradox for Field's theory in a language that has a fully determinate indeterminacy predicate. Field's theory avoids revenge paradoxes because it implies that the indeterminacy predicate in (3) displays indeterminacy. Of course, if one allows him this move, then he can justify his claim that his target language has all the resources needed to classify every type of revenge paradox. However, the sensible response to his claim is that the real revenge paradox concerns a sentence like (3) that contains an indeterminacy predicate that does not display any indeterminacy. If his theory applied to a language with such a predicate, then it would imply that (3) is both true and either false or indeterminate.[28] If one had access to exclusion negation, then one could easily define such an indeterminacy predicate. Of course, Field's response to this problem is to deny the intelligibility of a fully determinate indeterminacy predicate.[29] Thus, he has to appeal to the unintelligibility

[27] In Field (2003b: 302), he claims that an operator that could be used to define exclusion negation is "not fully intelligible."

[28] See Yablo (2003) for a similar criticism.

[29] Field (2003b: 302).

maneuver in two different ways to justify the claim that his theory is naturally internalizable. I have argued that if a theory appeals to the unintelligibility maneuver, then it is unacceptable.

Another reason to doubt that Field's theory is naturally internalizable is that his semantic theory for truth pertains to LS truth concepts. Of course, Field claims that natural language truth predicates can be explained in terms of LS truth predicates.[30] However, given the considerations in Appendix A, if his theory applies to natural languages then it is not descriptively correct, and if it does not, then it is not naturally internalizable.

In sum, Field's theory of truth marks a real advance over Kripke's because of the conditional that he defines. Moreover, his semantic theory for truth is admirable because it is both descriptively correct and internalizable for its target languages. Thus, his theory is weakly internalizable. In addition, it does not fall prey to self-refutation problems. However, it is not naturally internalizable.

D.3.5 Groundedness (Maudlin)

The final consistency theory I consider was published in 2004 by Maudlin.[31] Like those of Reinhardt, McGee, and Field, it is based on Kripke's theory of truth. Maudlin claims that his theory of truth does not require a substantive distinction between employed language and target language.

> If one has a formal language with the truth and falsity predicates, various grammatical predicates, the function $\mathcal{F}$(x), and a variable over functions from sentences in the language into the ordinals, the language can serve as its own metalanguage. Or at least, one can write down sentences in such a language which express the semantic theory which we have been considering, (Maudlin 2004: 86).

[30] Field (1994a, 1994b).

[31] Maudlin (2004). See Blamey (2002) for a similar theory.

Again, Maudlin's theory is complex and subtle, and I describe only enough of it to decide whether it is internalizable.

Maudlin's book contains many interesting and insightful discussions of truth and the liar paradox. Indeed, his emphasis on the inferential version of the paradox (i.e., the one that employs the truth rules instead of the one that employs the T-sentences) influenced my presentation of the paradox in Chapter One and my analysis of the revenge paradoxes and self-refutation problems in Chapter Three.[32] His theory of truth is very similar to Kripke's theory. Indeed, the principal differences are in his views on the employed language for the theory. Kripke claims that his theory can be formulated in a bivalent language, while Maudlin claims that the employed language should be thought of as gappy as well. In addition, Maudlin claims that all ungrounded sentences are gaps, which means that he endorses Kripke's minimal fixed point as the language that best describes the truth predicate (Kripke is noncommittal on this issue). One nice feature of Maudlin's theory is that it respects the truth rules. In order to do so, he has to offer a non-classical logic in which the inference rules of negation-introduction and conditional proof fail (though he does permit restricted versions of them).[33]

Is Maudlin's theory internalizable in any sense? Given that Maudlin's theory is a gap approach, a revenge paradox for it concerns:

(4) (4) is either false or a gap.

One consequence of Maudlin's theory is that the set of gappy sentences includes all the revenge liars and all the claims about the truth-statuses of the revenge liars (e.g., '(4) is true if and only if (4) is either false or a gap'), which were used in the derivation of the revenge paradoxes. Thus, Maudlin's theory does not face revenge paradoxes. As I argued in Chapter Three, theories of

[32] See Maudlin (2004: ch. 1).

[33] Maudlin (2004: ch. 6).

truth that respect the truth rules and avoid revenge paradoxes face self-refutation problems, and Maudlin's theory is no exception. For Maudlin, it is not just the consequences of his theory that are classified as gaps by his theory; indeed, his theory implies that most of the sentences that constitute it are gaps. He admits that on his theory of truth, his theory of truth is gappy. "The good news that one can write down the theory in the language is, however, matched by a piece of bad news. For the theory so expressed is, by its own lights, not true. That is, if one applies the method of semantic evaluation to the very sentences which express the semantic theory, those sentences mostly turn out to be ungrounded," (Maudlin 2004: 86).

In Chapter Three, I argued that theories of truth that respect the truth rules and face self-refutation problems are unacceptable because one can show that if a theory of truth respects the truth rules and implies that one of its consequences is a gap, then it is false (so long as one has access to a bivalent language with the requisite linguistic resources). Maudlin claims that there are no languages in which one could formulate this objection because the truth predicate for any language with the capacity to represent its syntactic structure is partial and the linguistic devices that could be introduced into such a language to formulate the objection (e.g., exclusion negation) are unintelligible.[34] Indeed, Maudlin claims that all non-monotonic sentential operators are unintelligible.[35] Thus, his response to the self-refutation problem constitutes an unintelligibility maneuver. (In Chapter Three, I called this the *robust response* to the self-refutation problem.)

He attempts to respond to the self-refutation problem by proposing an account of assertibility on which gappy sentences can be assertible. However, once one introduces the vocabulary into the target language to express this claim, one can generate a paradox that is very

[34] "No matter how much one wants there to be a Strong negation such that the Strong negation of an ungrounded sentence is true, such a connective is incoherent," (Maudlin 2004: 54).

[35] Maudlin (2004: 51). He also claims that strong truth is unintelligible; Maudlin (2004: 52).

much like the liar, but involves assertibility. Thus, it seems that Maudlin has traded in a paradoxical notion of truth for a paradoxical notion of assertibility. In response to this worry, he claims that assertibility is always relative to a set of rules and there is no ideal set of rules for assertibility. Given that the real problem with Maudlin's response to the self-refutation problem is that he is forced into making an unintelligibility maneuver, it seems to me that, although his claims about assertibility are implausible, an opponent should focus on the original self-refutation problem and the fact that his unintelligibility maneuver is unacceptable.

It seems to me that Maudlin does an excellent job of insuring that his theory of truth employs only the linguistic devices that belong to the target language. Thus, his semantic theory for truth is weakly internalizable and descriptively correct for its target language. Of course, his theory is also self-refuting, and so it is unacceptable, even if it is restricted to this target language. Because natural languages contain non-monotonic sentential operators, his semantic theory for truth is not naturally internalizable.

D.4 INCONSISTENCY THEORIES

All the theories of truth described in section three imply that truth is a consistent concept. I call these *consistency theories*. As I argued in Chapter Three, any consistency theory of truth that respects the truth rules faces either a revenge paradox or a self-refutation problem. As we saw, one can construct a weakly internalizable semantic theory for truth on the basis of a consistency theory of truth if one is willing to respond to either the revenge paradoxes or the self-refutation problems with an unintelligibility maneuver. Of course, such theories are radically implausible.

In this section, I consider four theories of truth that imply that truth is an inconsistent concept. I call these *inconsistency theories*. The major project for a proponent of an inconsistency theory of truth is constructing a plausible theory of inconsistent concepts. There has been little work done on this issue and there is little agreement on the nature of inconsistent concepts. Each theory of truth in this section is based on a different theory of inconsistent concepts.

In Chapter Three, I discussed a problem for any theory of inconsistent concepts; namely, most theories of X are *naïve theories* of X, which are collections of principles that purport to govern the concept X. However, if X is an inconsistent concept, then the principles that govern it are inconsistent; hence, if X is an inconsistent concept, then a naïve theory of X is an inconsistent theory. Thus, if inconsistent theories are unacceptable, then naïve theories of inconsistent concepts are unacceptable. If X is an inconsistent concept, then what form should a theory of X take? I have quite a bit to say in response to this question in Chapters Four, Five, Six, and Seven, and Appendix E. One important point for my purposes in this paper is that if X is an inconsistent concept, then an inconsistency theory of X should not employ X. Why? If X is an inconsistent concept, then sentences that express X are truth-value gaps (for the argument see Chapter Five).[36] Thus, if X is an inconsistent concept and a theory of X employs X, then some of the sentences that constitute that theory are gaps. On some accounts of assertibility, no such theory is assertible.[37] In addition, if X is an inconsistent concept, then X is a concept that should not be employed, not even in a theory of X. Presumably, once people stop using X, one should still be able to use the theory of X (otherwise, the explanation of X, the logic for X, and the semantic theory for X would have to be abandoned along with X). Thus, a theory of X

[36] This result holds for any concept of truth for which the truth rules are constitutive. It also holds for the replacement concepts of truth I offer in Chapter Seven. Thus, sentences that express inconsistent concepts are neither ascending weak true, ascending weak false, descending weak true, or descending weak false.
[37] See Appendix B for discussion.

should not employ X, if X is an inconsistent concept. Another reason for this claim is that when one employs an inconsistent concept, one commits oneself to following incompatible rules. Other things being equal, one should avoid committing oneself to following incompatible rules. Thus, one should not employ an inconsistent concept. That principle holds for theory construction as well. If one accepts a theory that employs an inconsistent concept, then one commits oneself to following incompatible rules. Thus, one should not accept any theory that employs an inconsistent concept.

These issues affect internalizability in the following way. If T is a semantic theory for truth and T is based on an inconsistency theory of truth that has the form of a naïve theory, then T is inconsistent unless it is restricted. The restriction will render T essentially external for natural languages. Likewise, if T is a semantic theory for truth, T is based on an inconsistency theory of truth, and T employs an inconsistent concept of truth, then T faces something similar to a revenge paradox unless T is restricted in a way that renders it essentially external. Consider how a revenge paradox affects someone who accepts a theory of truth. If S is a person who accepts a theory of truth that gives rise to a revenge paradox, then S is committed to the claim that the revenge liar is both true and not_E true. Now consider a person who accepts a theory that employs an inconsistent concept X. If S accepts such a theory, then S is employing X; hence, S is committed to the claim that X both applies and disapplies to the objects in X's overdetermination set. Thus, both the person who accepts a theory of truth that gives rise to a revenge paradox and the person who accepts a theory that employs an inconsistent concept commit themselves to a contradiction. A theory that employs an inconsistent concept can be restricted so that the items in the overdetermination set are outside its scope, but in the case of a semantic theory for truth that employs an inconsistent concept of truth, this restriction renders the theory essentially external for natural languages.

To summarize, an inconsistency theory of truth should not take the form of a naïve theory, and an inconsistency theory of truth should not employ the inconsistent concept of truth (or any other inconsistent concepts).

D.4.1 Dialetheism (Priest)

The first inconsistency theory of truth I consider is dialetheism. Actually, dialetheism is not, by itself, a theory of truth; rather, it is the view that some contradictions are true. The supporters of dialetheism use it together with a paraconsistent logic to construct a theory of truth. The logic used is called LP, and it is similar to the logic employed by Kripke's theory of truth, except that the sentences assigned truth-value gaps (neither true nor false) by Kripke's theory are assigned truth-value gluts (both true and false) by LP; on the logic at work in Kripke's theory, these sentences are not designated, while on LP, they are designated. LP is a *paraconsistent* logic, which means that it is not the case that everything follows from a contradiction in LP. However, LP is not trivial—it is not the case that all sentences are theorems. Thus, one can use it to draw a substantive distinction between theorems and non-theorems and between valid arguments and invalid arguments. With the help of LP and dialetheism, one can simply take the truth rules to be a theory of truth; call it the *dialetheic theory of truth*.[38]

[38] See Priest (1979, 1984, 1987, 1990, 1991, 1998, 2002). An axiomatic version of the dialetheic theory of truth adds the truth rules to the truths of first-order arithmetic or some axiomatizable theory of a fragment of first-order arithmetic to allow for self-reference; I ignore this complication.

It is important to understand something about dialetheism. It is not the view that some contradictions are rationally acceptable. Field endorses this interpretation and that is a mistake.[39] There is a cognitive/aletheic ambiguity that haunts areas around here. The problem seems to be that 'true' can be used to convey agreement or acceptance. However, this cannot be the meaning of 'true' because it fails in embedding tests. Nevertheless, it is easy to confuse the illocutionary aspect of truth with its locutionary aspect. That is, it is easy to mistake a claim that p is true for the claim that one believes that p. Field makes this mistake when he revises the account of dialetheism in terms of belief or acceptance. Dialetheism, as it as been formulated and defended, is an aletheic doctrine, not a cognitive one. One can imagine a doctrine, *dicognitivism*, which implies that it is rational for us to believe contradictions, or *diconativism*, which implies that it is rational for us to desire contradictory circumstances, or *diassertionism*, which implies that it is rational to assert a contradiction. However, these are distinct from dialetheism, which is a doctrine about the features of truth bearers.

Notice that the dialetheic theory of truth is a naïve theory of truth. It is also inconsistent. The proponents of dialetheism admit that this theory is inconsistent (indeed, they claim that it is both true and false), and they have spent a considerable amount of time and energy arguing that inconsistent theories can be rationally acceptable.[40] It seems to me that the principle of mono-aletheism (i.e., no truth-bearer is both true and false) is constitutive of our concept of truth. Thus, it seems to me that one can reject dialetheism and the dialetheic theory of truth without giving any further justification.[41] However, in this paper, I am interested in whether the theories I consider are internalizable in any of the three senses.

[39] Field (forthcoming a).

[40] See Chihara (1984), Priest (1984, 1987, 1998, 2000), Parsons (1990), and Field (2002).

[41] See Lewis (1982: 101) for a similar view.

I am willing to assume that the dialetheic theory of truth appeals only to concepts definable in LP. Thus, the semantic theory for truth that is based on this theory is weakly internalizable. Of course, it is not descriptively correct (because it is inconsistent), but the dialetheist will deny that this is a problem and I am not going to argue the point here. The semantic theory for truth that is based on the dialetheic theory is not naturally internalizable because of a familiar problem. The dialetheic theory cannot be applied to languages with non-monotonic sentential operators because LP is trivial for such languages; that is, every sentence of such a language is a theorem. Even the dialetheist cannot accept a trivial logic. Thus, in order to keep his theory dialetheically acceptable, the dialetheist must restrict it to languages with no non-monotonic sentential operators. Proponents of the dialetheic theory have responded to this criticism by denying that non-monotonic sentential operators are intelligible.[42] Thus, the dialetheist too appeals to an unintelligibility maneuver to justify using his theory on natural languages.

Before moving on, I want to note that the dialetheic theory of truth is the result of applying the dialetheic theory of inconsistent concepts to truth. On the dialetheic theory of inconsistent concepts, an acceptable theory of an inconsistent concept X is a naïve theory of X. Of course, a naïve theory of X will be inconsistent (and both true and false according to the dialetheist), but provided one accepts LP and the claim that some inconsistent theories are rationally acceptable, one might not see this as a problem. The fact that the dialetheic theory of truth is unacceptable because it is not naturally internalizable is significant because it shows that even if one is willing to give up the law of non-contradiction in an attempt to solve the liar paradox, one's theory is still unacceptable. The problem, as I see it, is that although the dialetheist is on the right track in accepting that truth is an inconsistent concept, he still

[42] Priest (1990; 2002: 384-5).

constructs a naïve theory of truth and his theory of truth employs the inconsistent concept of truth. It is essential to recognize that one can hold that some concepts are inconsistent, but reject dialetheism.

D.4.2 Inconsistent Definitions (Yablo)

Yablo presents a different account of inconsistent concepts by way of a theory of inconsistent definitions. He then uses this theory to arrive at a theory of strong truth. A truth predicate is *strong* if and only if 'p is true' is false if p is a gap (see Chapter Seven for discussion). Although Yablo does not discuss issues associated with internalizability, his theory has some interesting features.[43]

Let the canonical form of a definition be: $Px =_{\text{def}} \phi(x)$, where $\phi(x)$ is a formula with only 'x' free. He begins by distinguishing between noncircular, positive circular, and negative circular definitions. A *noncircular definition* is one where P does not occur in ϕ, a *positive circular definition* is one where P occurs in φ in such a way that if P's extension increases, ϕ's does as well, and a *negative circular definition* is one where P occurs in ϕ in such a way that if P's extension increases, ϕ's decreases. As for noncircular definitions, we can follow rule (D) to determine their extension on the basis of the definition:

(D) x satisfies P in world w if and only if x belongs to ϕ's extension in w.

To use (D) one must first determine ϕ's extension in w, which I will assume can be done without trouble for noncircular definitions.

[43] Yablo (1985, 1993a, 1993b).

Rule (D) will not work for circular definitions because they require one to first know the extension of P in order to determine the extension of ϕ. One might instead follow rule (E) to determine the extension of P:

> (E) x satisfies P if and only if x is a member of Φ, where Φ is a set that solves the equation $\ulcorner \Phi = \phi_{w} (\Phi) \urcorner$ and $\phi_{w}(\Phi)$ is the extension of ϕ in w on the assumption that P has Φ as its extension.

If Φ solves the above equation, then setting P's extension to Φ will make the definition come out true. One problem for using (E) as a rule for determining the extension of circularly defined expressions is that the equation associated with a positive circular definition will usually have multiple solutions and that associated with a negative circular definition need not have any solution at all.[44]

To solve this problem, one can rely on rule (F) for positive circular definitions:

> (F) x satisfies P in w if and only if x is a member of Φ, where Φ is a set that constitutes the *least* solution to the equation $\ulcorner \Phi = \phi_{w} (\Phi) \urcorner$ and $\phi_{w}(\Phi)$ is the extension of ϕ in w on the assumption that P has Φ as its extension.

The difference between (E) and (F) is that the latter forces the extension of F to be the least solution to the equation associated with F's definition and one can prove that any such equation has a least solution.

For negative circular definitions, one can use:

> (G) x satisfies P in w if and only if x is Δ-grounded in w.

Unfortunately, the definition of 'Δ-grounded' is a bit complicated because it incorporates Tarski's definition of satisfaction. The following are the rules that are based on Tarski's definition:

> (AT) $\mathbf{s}(x) \in \mathbf{A} \rightarrow T(Ax, \mathbf{s})$.

[44] Positive circular definitions are also called inductive definitions, negative circular definitions are also called anti-inductive definitions. See Yablo (1985; 1993a: appendix).

(AF) $\mathbf{s}(x) \in \mathbf{A} \rightarrow F(Ax, \mathbf{s})$.

(~T) $F(\psi, \mathbf{s}) \rightarrow T({\sim}\psi, \mathbf{s})$.

(~F) $T(\psi, \mathbf{s}) \rightarrow F({\sim}\psi, \mathbf{s})$.

($\wedge$T) $T(\psi, \mathbf{s})$ and $T(\chi, \mathbf{s}) \rightarrow T(\psi \wedge \chi, \mathbf{s})$.

($\wedge$F) $F(\psi, \mathbf{s})$ and $F(\chi, \mathbf{s}) \rightarrow F(\psi \wedge \chi, \mathbf{s})$.

($\forall$T) $T(\psi, \mathbf{s}')$ for all $\mathbf{s}' \approx_x \mathbf{s} \rightarrow T(\forall x \psi, \mathbf{s})$.

($\forall$F) $F(\psi, \mathbf{s}')$ for all $\mathbf{s}' \approx_x \mathbf{s} \rightarrow F(\forall x \psi, \mathbf{s})$.

$\ulcorner$T (ψ, **s**)$\urcorner$ is synonymous with $\ulcorner$formula ψ is true of the object assigned to its variable by the function, **s**, which assigns objects from the domain to the variables of the language$\urcorner$, and $\ulcorner$F (ψ, **s**)$\urcorner$ is synonymous with $\ulcorner$formula ψ is false of the object assigned to its variable by the function, **s**, which assigns objects from the domain to the variables of the language$\urcorner$. Rest assured, these are just the standard clauses from Tarski's definition of satisfaction. I refer to these as the *satisfaction rules*. The rest of the rules are:

(ΔT) $T(\phi, \mathbf{x}) \rightarrow T(P, \mathbf{x})$.

(ΔF) $F(\phi, \mathbf{x}) \rightarrow F(P, \mathbf{x})$.

The first says that one should add x to P's extension if the satisfaction rules prove that ϕ is true of it; the second says that one should not add x to F's extension if the satisfaction rules prove that ϕ is false of it. Now we can define Δ-groundedness:

(Δ-groundedness) an object is Δ-grounded if and only if the satisfaction rules, (ΔT), and (ΔF) prove $T(\phi, \mathbf{x})$.

It turns out that one can rephrase (E) in terms of proof by the satisfaction rules and one can rephrase (F) in terms of the satisfaction rules and (ΔT). Thus, (G) will work for non-circular, positive circular, and negative circular definitions.

There is one more set of definitions that have yet to be dealt with. These definitions are circular but are neither positive nor negative because the definiendum occurs both positively and negatively in the definiens. For these, Yablo suggests a reflection rule that incorporates the reasoning from the groundlessness of the claim that ϕ is true of x to the claim that P is false of x:

(ΔR) $\Theta \to F(P, x)$, where Θ makes x ungroundable.

Θ is a set of claims about which objects satisfy which formulas and Θ makes x ungroundable if and only if $T(\phi, x)$ is not provable using the satisfaction rules and (ΔT), from the set of $F(P, s)$ such that $T (P, s) \notin \Theta$. Rule (ΔR) allows us to infer that some object is not in P's extension from the fact that $T(\phi, x)$ is not provable from the satisfaction rules and (ΔT) (i.e., it allows us to infer that x is not in P's extension from the fact that $T(\phi, x)$ is groundless).

Finally we get to inconsistent definitions. For Yablo, a definition is *consistent* if and only if (E), (F), and (G) are jointly satisfiable. A definition is *inconsistent* if and only if it is not consistent. As a test of consistency, Yablo defines two sets:

Γ_{Δ} = {x: the reflective rules prove T(P, x)}

Γ^{Δ} = {x: the reflective rules do not prove F(P, x)}

Here, the *reflective rules* are the satisfaction rules, (ΔT), and (ΔR). A definition is consistent if and only if $\Gamma_{\Delta} = \Gamma^{\Delta}$. Any attempt to follow the semantic rules when dealing with inconsistent definitions is impossible. Yablo draws an analogy with incompatible moral obligations; the difference is that, with inconsistent definitions, an attempt to comply with one obligation creates another that one must defy.[45]

Yablo's theory of truth results from applying this account of inconsistent definitions to an inconsistent definition of truth. Yablo endorses the following definition of truth:

[45] Yablo (1993a, 1993b).

ϕ is true $=_{df}$ either $\phi = \ulcorner Ra \urcorner$ and a’s referent belongs to R’s extension, or
$\phi = \ulcorner {\sim}\psi \urcorner$ and ψ is false, or
$\phi = \ulcorner \psi \;\&\; \chi \urcorner$ and both ψ and χ are true, or
$\phi = \ulcorner \forall x \psi(x) \urcorner$ and all its instances are true, or
$\phi = \ulcorner \psi$ is true$\urcorner$ and ψ is true.

ϕ is false $=_{df}$ either $\phi = \ulcorner Ra \urcorner$ and a’s referent does not belong to R’s extension, or
$\phi = \ulcorner {\sim}\psi \urcorner$ and ψ is true, or
$\phi = \ulcorner \psi \;\&\; \chi \urcorner$ and either ψ or χ are false, or
$\phi = \ulcorner \forall x \psi(x) \urcorner$ and some of its instances are false, or
$\phi = \ulcorner \psi$ is true$\urcorner$ and ψ is not true.

This definition is circular, but it is neither positive nor negative; it is inconsistent. The notion of truth defined here is known as *strong truth* (see Chapter Seven for discussion). Yablo uses his account of inconsistent definitions to arrive at a theory of strong truth and a semantic theory for strong truth.

Is Yablo’s theory internalizable in any sense? It seems to me that his theory of inconsistent definitions does not require any concepts that are inexpressible in the target language of his theory. Thus, the semantic theory for strong truth that is based on his theory of strong truth is weakly internalizable. Notice that it is not a naïve theory. However, Yablo’s theory of strong truth does employ the inconsistent concept of strong truth it purports to describe. Thus, it faces something like a revenge paradox. Anyone who accepts Yablo’s theory of strong truth is committed to following incompatible rules for employing the concept of strong truth. Granted, according to Yablo, although the constitutive principles for strong truth are inconsistent, his theory of truth implies that these principles are not all in effect simultaneously. The rules for using it change from context to context. In some contexts the rules stipulate that it should be employed in one way and in a different context, they stipulate that it should be employed in a different way. Indeed, following its rules in a given context can change the context in a way that also changes its rules. One can find oneself employing the concept of strong truth in a context where one follows its rules and applies it to a sentence, p. This application changes the context and changes the rules so that one should now disapply it to the same object, sentence p. This employment again changes the context, and in the new context, one should now apply it to p. And so on, forever. Anyone who accepts Yablo's theory employs the strong concept of truth and so employs an inconsistent concept. However, anyone who employs it accepts an inconsistent set of principles and is committed to flip-flop like this on certain sentences.

Furthermore, Yablo's theory faces something like a self-refutation problem as well. Some of the sentences on which one flip-flops are consequences of the theory. Thus, acceptance of the theory requires employment of the strong concept of truth, which results in flip-flopping on some consequences of the theory and so on the theory itself.

One version of Yablo's theory applies to natural languages and is naturally internalizable; however, this version of the theory is unacceptable because of the flip-flop problems. If Yablo's theory is to be acceptable, then it must be restricted so that sentences that initiate flip-flop problems are outside its scope. Obviously, this restriction renders the theory essentially external for natural languages.

D.4.3 Supervaluation (Eklund)

In this section, I present Eklund's theory of inconsistent concepts and his inconsistency theory of truth. Both appeal to a logic device called supervaluation (see Appendix E for discussion). Eklund does not discuss issues associated with internalizability, but as I argue, his theory fares relatively well when compared to the others I have discussed.

Eklund provides a theory of truth and a theory of vague concepts on which truth and vague concepts are inconsistent concepts.[46] He argues that by virtue of our semantic competence, we accept the premises and the inference rules that lead to the liar paradox and the sorites paradox. Eklund phrases his analysis in terms of inconsistent languages, but I prefer to concentrate on concepts because of the flexibility this allows. One can think of an inconsistent language as one that expresses an inconsistent concept. For Eklund, a concept is inconsistent if and only if the set of constitutive principles for it is inconsistent, and one must accept all the constitutive principles for a concept in order to possess that concept. He resists the temptation to think of constitutive principles as true or unrevisable. To clarify this claim, he introduces the notions of competence dispositions and culprits. One's *competence dispositions* are belief-forming dispositions that one has by virtue of one's semantic competence. A *culprit* is the false premise or invalid inference used in the derivation of the contradiction in a paradox. One can say that one's competence dispositions lead one to accept the culprit of the paradox because the set of cognitive meaning-constitutive sentences associated with the concept in question is inconsistent. If a concept displays this phenomenon, then the paradox associated with it is said to *exert pull*. One of Eklund's central theses is that the liar paradox and the sorites paradox exert pull (I call this the *pull exertion thesis*).

Eklund considers the compatibility of the pull exertion thesis with three popular theories of semantic competence: the truth conditional theory, conceptual role semantics, and the Fregean theory. For the first, to know the meaning of a sentence is to know its truth conditions. On conceptual role semantics, to know the meaning of a sentence is to have the right belief-forming dispositions. Fregean theories imply that the semantic values of sentences satisfy their senses. He argues that all three preclude the pull exertion thesis. The first would require an inconsistent

[46] Eklund (2002).

set of truths, the second would require an unsound set of valid inference rules, and the third would require semantic values to satisfy an unsatisfiable set of conditions. If the liar paradox and the sorites paradox do exert pull then there must be something wrong with these three accounts of semantic competence (see Appendix E for discussion of a similar argument due to Gupta).

To remedy this situation, Eklund suggests some changes for conceptual role theories and for Fregean theories. The remedy for the former is to allow competence dispositions for the acceptance of inferences that are not truth preserving and that are defeasible. Fregean theories need to allow the sense of an expression to be the constitutive principles associated with it. The sense of an expression then determines that the semantic value of an expression is whatever comes closest to satisfying these principles.

Eklund's suggestion for a semantics for inconsistent concepts is to define an acceptable assignment of semantic values to expressions of an inconsistent language, L, as one that makes true a weighted majority of the constitutive principles for L. Eklund says very little on how to determine whether an assignment is acceptable and on the weighting function that should be used. In a footnote, Eklund compares his strategy with the traditional supervaluation semantics:

> Talk of "acceptable assignments" is familiar from supervaluationist analyses. But note that the acceptable assignments considered here are quite different from the acceptable assignments—SV-assignments, let us call them—the supervaluationist talks about. For the supervaluationist, the reason there are many SV-assignments for a natural language is that the meanings of some expressions are incomplete: they can be extended without being changed, as Kit Fine puts it (1975, p. 267). The SV-assignments correspond to all the possible completions of meanings of expressions of the language. They are not, as the acceptable assignments discussed here, meant to be as faithful as possible to the meanings expressions actually are endowed with. (As illustrated by, for example, the fact that although supervaluationists normally do not accept bivalence, all the particular SV-assignments are bivalent), (Eklund 2002, 265n33).

As I understand it, the big difference (according to Eklund) between his theory and supervaluation semantics is that, for supervaluation semantics, the semantic values of the expressions in question are determined by considering a collection of assignments and constructing one on the basis of their shared properties, whereas Eklund's theory determines semantic values by considering a collection of assignments and picking one (or more) from among them on the basis of which ones satisfy the constitutive principles associated with the expressions of the language. On the former, an expression has a certain semantic feature if all the various acceptable assignments imply that it does; on the latter, an expression has a certain semantic feature if the claim that it does is the most compatible with the constitutive principles.

It seems to me that this way of putting the difference is misleading. The real difference is just that most supervaluation semantics do not respect the constitutive principles associated with the expressions of a language. That is, a supervaluation semantics treats all constitutive principles as aletheic—i.e., ones that must be true. Eklund places a further constraint on what counts as an acceptable assignment. An acceptable assignment must make a weighted majority of the constitutive principles associated with the expressions of a language come out true. In (still) other words, Eklund's version of supervaluation semantics allows sentences that express penumbral connections to turn out false.[47] Another difference seems to be that Eklund does not take a stand on what to do in the face of multiple incompatible acceptable assignments. He suggests that it might be all right to retain bivalence at the cost of accepting multiple incompatible assignments for some expressions.[48]

[47] Tappenden (1993) makes similar recommendations and Fodor and Lepore (1996) point out the tension between constitutive principles and supervaluation semantics. I certainly do not agree with Fodor and Lepore's conclusion or their argument strategy, but the tension they point out is a real one; see also Morreau (1999).

[48] Eklund (2002: 263-6).

Let us consider the theory of truth and the semantic theory for truth that are based on Eklund's theory of inconsistent concepts. I assume that the truth rules (i.e., *ascending*: from ⟨p⟩ infer ⟨⟨p⟩ is true⟩; *descending*: from ⟨⟨p⟩ is true⟩ infer ⟨p⟩; and *substitution*: any name for ⟨p⟩ can be substituted in ⟨⟨p⟩ is true⟩ without changing its truth-status) and the rule of *mono-aletheism* (i.e., no truth bearer is both true and false) are constitutive principles for truth; these principles are inconsistent. The liar paradox results from the fact that it follows from these principles that the liar sentence is both true and not_{E} true.[49] The liar paradox exerts pull because by virtue of our competence dispositions, we are led to accept the culprits in the derivation of the contradiction.

Eklund's semantic theory for truth implies that an acceptable assignment of semantic values to the sentences that contain truth predicates will be one that makes true a weighted majority of the constitutive principles for truth. Which ones should be privileged? It seems to me that substitution and mono-aletheism are not negotiable.[50] That leaves ascending and descending. Which one is more important? It seems to me that descending is more important, but that intuition is probably idiosyncratic. Either way, any acceptable assignment of semantic values to truth sentences makes true substitution, mono-aletheism, and either ascending or descending. Presumably, the untrue truth rule is assigned either falsity or gaphood (depending on how one sets up the theory). It seems to me that this theory has a host of problems (which I discuss in Appendix E), but here I am concerned with internalizability.

Is Eklund's theory internalizable? Before answering this question, one should note that his theory is not a naïve theory of truth, but it does employ the inconsistent concept of truth. Thus, it satisfies one of the constraints presented at the beginning of section three, but not the

[49] 'not_{E}' expresses exclusion negation.

[50] See Eklund (2002: 267).

other. Given that the acceptable assignments of truth values must satisfy the law of mono-aletheism, and the theory does not respect all the truth rules, Eklund's theory does not give rise to revenge paradoxes. Of course, he explains why they exert pull and such. His theory says the same things about revenge paradoxes that it says about the liar paradox; if the liar is false (gappy) on Eklund's theory then so is the revenge liar. Thus, he might need a non-classical logic to accompany his theory, but I am willing to assume that he can do this without appealing to concepts outside his target language. Thus, his theory is weakly internalizable, and it is descriptively correct for the target language (given that the target language is a formal language with stipulated semantic features).

Does the theory face self-refutation problems? Yes. Assume that the theory implies that the liar is false. Then, 'the liar is false' is a consequence of the theory. Thus, the liar is a consequence of the theory. Thus, the theory implies that one of its consequences is false. If, instead, it assigns the liar gaphood, then it assigns the revenge liar for gap approaches a gap as well. Thus, 'the revenge liar is either false or gappy' is a consequence of the theory. Hence, the revenge liar is a consequence of the theory. Therefore, the theory implies that one of its consequences is a gap. Either way, the theory is self-refuting. Because it does not validate the truth rules, he might be able to get out of this problem by proposing an account of assertibility and rational acceptability on which gappy sentences are assertible and acceptable. If so, then he will face the assertibility paradox that Maudlin faces.

Because his theory employs the inconsistent concept of truth it purports to describe, it does face an analog of the revenge paradox (the problem is similar to the one confronting Yablo's theory). In particular, if S accepts Eklund's theory, then S employs the inconsistent concept of truth. Thus, if one accepts Eklund's theory, then one is disposed to accept the truth rules. If one accepts the truth rules, then one is disposed to accept that paradoxical sentences are both true and not_E true. I showed in the previous paragraph that a paradoxical sentence is a consequence of Eklund's theory. Thus, anyone who accepts Eklund's theory is disposed to accept that it has a contradiction as a consequence. Although his theory can be formulated so that it does not imply that it is false, it does imply that anyone who accepts it will be disposed to accept that it is false. This feature of his theory stems from the fact that his theory employs the inconsistent concept of truth. It seems to me that any theory with this feature is unacceptable. Nevertheless, a version of his theory is naturally internalizable. However, it is both self-refuting and unacceptable because it employs an inconsistent concept. One could construct a restricted version of Eklund's theory that avoids both the self-refutation problem and the analog of the revenge paradox, but the restricted version is not naturally internalizable.

D.4.4 Confusion

In this final section, I discuss the theory of inconsistent concepts I endorse. On this theory, an inconsistent concept is one whose constitutive principles are inconsistent. The central claim of this theory is that inconsistent concepts are confused concepts. That is, for each inconsistent concept, there is a set of component concepts that play two roles. First, they are used in the logic, the pragmatic theory, and the semantic theory for inconsistent concepts; second, they serve as replacements for the inconsistent concept.

When considering an inconsistent concept, it is essential to distinguish between several sets of rules for using it. First, there are the inconsistent rules that are constitutive of the concept. Those who employ the concept try to follow these rules. Second, there are the rules stipulated by the logic, the pragmatic theory, and the semantic theory for inconsistent concepts. An interpreter who knows the concept is inconsistent treats those who employ it as if they are bound by these rules. Third, there are the rules stipulating that the concept should not be used at all. Those who know that it is inconsistent are bound by these rules.

It is essential to distinguish between an inconsistent concept's application set, its extension, its disapplication set, its anti-extension, its range of inapplicability, and its non-extension. A concept's *application set* includes all the items to which applies, its *disapplication set* contains all the items to which it disapplies, and its *range of inapplicability* consists of all the items to which it neither applies nor disapplies (all three are determined by its constitutive principles). (The union of its application set and disapplication set is its *range of applicability*.) For acceptable concepts, the extension and the application set are identical, the anti-extension and the disapplication set are identical, and the non-extension and the range of inapplicability are identical. An inconsistent concept has an empty extension and anti-extension, but its application set and a disapplication set need not be. An inconsistent concept can be application-inconsistent or range-inconsistent (or both). A concept is *application-inconsistent* if and only if its application set and disapplication set are not disjoint. A concept is *range-inconsistent* if and only if its range of applicability and range of inapplicability are not disjoint.

The theory of inconsistent concepts I offer has a logic, a pragmatic theory, and a semantic theory for inconsistent concepts. The logic appropriate for an inconsistent concept depends on its components. If it is completely defined and has n components, then an n-component logic is appropriate. If it is partially defined and has n components, then a partial n-component logic is appropriate. Both n-component logics and partial n-component logics are relevance logics. The pragmatic theory is a scorekeeping theory—it specifies how those who know the concept is inconsistent should keep score on those who employ it. The semantic theory for inconsistent concepts is an inferential role theory—it specifies the inferential roles of the sentences that express the inconsistent concept (the inferential role of a sentence includes its role in perception and action as well). These three theories are used to interpret those who employ inconsistent concepts. In addition, I endorse the replacement policy for handling cases of conceptual inconsistency; inconsistent concepts should be replaced with consistent ones.

When one uses this theory of inconsistent concepts to arrive at an inconsistency theory of truth, one must decide on the components of truth. My view is that truth is both range-inconsistent and application-inconsistent. One can deal with the range-inconsistency by distinguishing between weak truth, strong truth, and dual truth, but each of these concepts is still application-inconsistent (see Chapter Seven for discussion of these concepts). One can deal with the application-inconsistency by distinguishing an ascending and a descending version of each of these concepts. The theory of truth I offer implies that our inconsistent concept of truth has six components: weak ascending truth, weak descending truth, strong ascending truth, strong descending truth, dual ascending truth, and dual descending truth. The strong and dual concepts are defined in terms of the weak concepts.

The weak truth predicates are defined in the following way. Both weak truth predicates are partially defined and have the same range of applicability; however their extensions and anti-extensions are slightly different. Ascending weak truth obeys the rule: from ⟨p⟩ infer ⟨⟨p⟩ is ascending weak true⟩. Descending weak truth obeys the rule: from ⟨⟨p⟩ is descending weak true⟩ infer ⟨p⟩. Ascending weak truth obeys the analog of the descending weak truth rule for non-pathological sentences, and descending weak truth obeys the analog of the ascending weak truth rule for non-pathological sentences. A *pathological* sentence is a sentence that contains either 'ascending weak true' or 'descending weak true' and would be weak paradoxical if they were replaced with the inconsistent weak truth predicate; a sentence is *weak paradoxical* if and only if the weak truth rules imply that it is both weak true and not$_E$ weak true. Ascending weak truth and descending weak truth are weak truth predicates because if ⟨p⟩ is a weak gap, then ⟨⟨p⟩ is ascending weak true⟩ and ⟨⟨p⟩ is descending weak true⟩ are both weak gaps. All the sentences that expresses defective concepts (including all the sentences that express our inconsistent concept of truth) are weak gaps. It is important to recognize that although ascending and descending weak truth are partial, they are consistent concepts.

Ascending weak truth and descending weak truth differ on the pathological sentences. Consider the following two sentences:

(α) (α) is ascending weak false.

(δ) (δ) is descending weak false.

These sentences are pathological. However, they are not paradoxical—it is not the case that the theory of the weak truth predicates has the following consequences: (i) (α) is both ascending weak true and ascending weak false, (ii) (α) is both descending weak true and descending weak false, (iii) (δ) is both ascending weak true and ascending weak false, and (iv) (δ) is both descending weak true and descending weak false. Indeed, the theory of the weak truth predicates implies that (α) and (δ) are both ascending weak true and descending weak false. Consider the analogs of the liar reasoning for (α) and (δ):

(α) is AWT (assumption)	(δ) is DWT (assumption)
'(α) is AWF' is AWT (substitution)	'(δ) is DWF' is DWT (substitution)
(α) is AWF (descending)	(δ) is DWF (descending)
(α) is AWF (assumption)	(δ) is DWF (assumption)
'(α) is AWF' is AWT (ascending)	'(δ) is DWF' is DWT (ascending)

(α) is AWT (substitution)	(δ) is DWT (substitution)
∴(α) is AWT iff (α) is AWF	∴(δ) is DWT iff (δ) is DWF

Neither of these arguments is valid. In the argument concerning (α), the third step is invalid, and in the argument concerning (δ), the fifth step is invalid. Of course, one can prove that both (α) and (δ) are AWT and DWF, but that is not a contradictory conclusion.

Figure D.1 is a diagram of weak truth (which is inconsistent) and Figure D.2 is a diagram of ascending weak truth and descending weak truth (which are consistent).

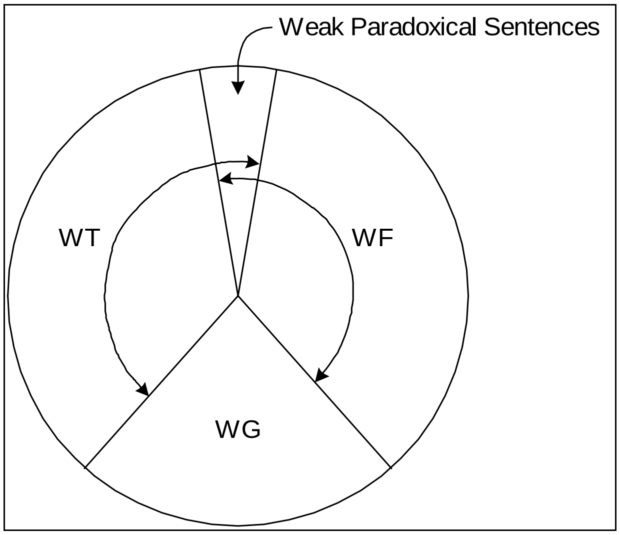


Figure D.1 (Weak Truth)

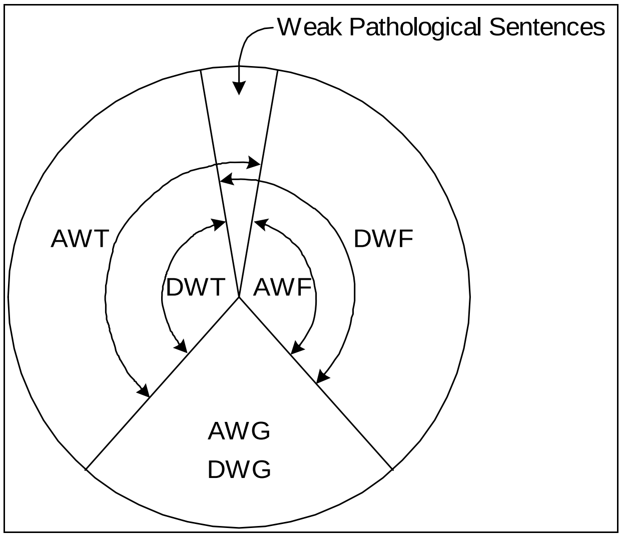


Figure D.2 (Ascending Weak Truth and Descending Weak Truth)

Notice that Figure D.1 is a diagram of the application set, disapplication set, and range of inapplicability for weak truth, while figure D.2 is a diagram of the extension, anti-extension, and non-extension of ascending weak truth and descending weak truth (weak truth has an empty extension and an empty anti-extension).

The components of truth are used in the logic, the pragmatic theory, and the semantic theory for truth. Because truth has six components and they are partially defined with different ranges of applicability, the logic for truth is a partial 6-component logic. This logic has 63 epistemically-interpreted semantic values (see Chapters Six and Seven for some details). The partial 6-component logic is a relevance logic—it does not validate disjunctive syllogism, and it does not imply that every sentence is a consequence of a contradiction.

The liar reasoning is invalid in the partial 6-component logic. Consider the liar reasoning (let (λ) be '(λ) is false'):

(a) (λ) is true. (assumption)

(b) '(λ) is false' is true. (from (a) by substitution)

(c) (λ) is false. (from (b) by descending)

(d) if (λ) is true, then (λ) is false. (from (a) through (c) by conditional proof)

(e) (λ) is false. (assumption)

(f) '(λ) is false' is true. (from (e) by ascending)

(g) (λ) is true. (from (f) by substitution)

(h) if (λ) is false, then (λ) is true. (from (e) through (g) by conditional proof)

(i) (λ) is true and (λ) is false. (from (d) and (h) by classical logic)

Steps (c) and (f) are invalid. (Argument: the AW query value for (a) is Y, for (b) is Y, and for (c) is N; the DW query value for (a) is Y, for (b) is Y, and for (c) is Y. Thus, step (c) is invalid. The AW query value for (e) is Y, for (f) is Y, and for (g) is Y; the DW query value for (e) is Y, for (f) is N, and for (g) is N. Thus, step (f) is invalid.) In general, any argument that depends on applying one of the truth rules to a paradoxical sentence is invalid in the logic for truth.

Is the semantic theory for truth internalizable in any of the three ways? To answer this question, we need to decide whether the inconsistency theory of truth faces any self-refutation problems or revenge paradoxes. I cannot guarantee that it does not. However, given that my analyses of self-refutation problems and revenge paradoxes are correct, I can argue that it does not. First, notice that the inconsistency theory of truth is not a naïve theory. It does not imply that the constitutive principles for truth are true or valid. In addition, this theory of truth does not employ the inconsistent concept of truth. The logic does not use truth-values and it does not explain validity in terms of truth preservation; instead, it uses epistemically interpreted semantic values and it explains validity in terms of profitability preservation. The pragmatic theory does not explain assertibility in terms of truth; instead, it uses the same resources as the logic. The semantic theory does not assign truth conditions to the sentences in its scope; instead, it assigns inferential roles. Thus, one can accept this theory of truth without employing the inconsistent concept of truth it describes.

My explanation of the revenge paradoxes is that they occur for theories of truth that respect the truth rules and classify some of the paradoxical sentences as defective, where the class of defective sentences does not include all the paradoxical ones. However, the inconsistency theory of truth I endorse classifies all the sentences that express the inconsistent concept of truth as defective. Thus, it does not face any revenge paradoxes. The theory of the component concepts of truth does not respect the truth rules; hence it does not give rise to revenge paradoxes.

Self-refutation problems occur for theories of truth that respect the truth rules and classify all the paradoxical sentences as defective, including some of their own consequences. However, the inconsistency theory of truth has no paradoxical sentences as consequences because it does not employ the inconsistent concept of truth. It does not classify the sentences in its scope as true or false. Instead, it implies that all sentences are in the range of inapplicability for the inconsistent concept of truth. The theory of the component concepts classifies all the sentences that express the inconsistent concept of truth as weak gaps. Still, one might wonder whether the theory of ascending weak truth or the theory of descending weak truth faces a self-refutation problem. Call the theory of ascending weak truth T_{AW} and the theory of descending weak truth T_{DW}. It might seem that both theories are pathological (i.e., AWT and DWF). Even if this were descending weak true, it would not constitute a self-refutation problem because pathological sentences are acceptable. However, neither theory is pathological. Consider T_{AW}. It implies that (α) is AWT and DWF. One might be tempted to infer from the claim that T_{AW} implies that (α) is AWT, that T_{AW} implies (α). However, the rule, from ⟨⟨p⟩ is AWT⟩ infer ⟨p⟩, is invalid for pathological sentences. Now consider T_{DW} and the following argument:

T_{DW} implies that (δ) is AWT and DWF.

(δ) is '(δ) is DWF'.

Hence, (δ) is a consequence of T_{DW}.

Therefore, T_{DW} is AWT and DWF.

The problem with this argument is that pathologicality is not preserved by the consequence relation. One can define validity in terms of AWT and DWT: an argument is *valid* if it preserves DWT and the absence of AWF (i.e., if the premises are DWT, then the conclusion is DWT, and if the conclusion is AWF, then one of the premises is AWF). '(δ) is DWF' is a consequence of T_{DW}, and '(δ) is DWF' is (δ); hence, (δ) is a consequence of T_{DW}. However, in order to show that T_{DW} is pathological, one would have to show that '(δ) is DWT' is a consequence of T_{DW}.

However, '(δ) is DWT' is not a consequence of T_{DW} because '(δ) is DWF' is a consequence of T_{DW}. Therefore, neither T_{AW} nor T_{DW} is pathological.

To summarize, the inconsistency theory of truth I offer does not give rise to revenge paradoxes, it is not self-refuting, and it does not employ the inconsistent concept of truth. Moreover, the theories of the component concepts of truth do not give rise to revenge paradoxes, they are not self-refuting, and they do not employ the inconsistent concept of truth. In addition, the component concepts are consistent concepts. Furthermore, I do not appeal to an unintelligibility maneuver to secure these results. The inconsistency theory of truth is not restricted from applying to the language in which it is formulated; hence, the semantic theory for truth that is based on it is weakly internalizable. The inconsistency theory of truth is not restricted from applying to natural languages; hence, the semantic theory for truth that is based on it is naturally internalizable. Moreover, the semantic theory for truth is descriptively correct—it is consistent and it assigns the correct meanings to the sentences in its scope. Finally, the inconsistency theory of truth, the logic for truth, the pragmatic theory for truth, the semantic theory for truth, and the theory of the component concepts are all descending weak true. Therefore, the inconsistency theory of truth I offer is the only one of the theories I have discussed that is acceptable for use on a natural language.

D.5 CONCLUSION

In section two, I presented an account of internalizable semantic theories and three internalizability requirements. I then discussed nine purportedly internalizable semantic theories for truth: Reinhardt's theory, McGee's theory, Simmons' theory, Field's theory, Maudlin's theory, the dialetheic theory, Yablo’s theory, Eklund’s theory, and the theory I propose. I concluded that Simmons’ theory is not weakly internalizable. Reinhardt’s theory, McGee’s theory, Field’s theory, Maudlin’s theory, and the dialetheic theory are weakly internalizable, but not naturally internalizable. Field’s theory, Maudlin’s theory, and the dialetheic theory appeal to an unintelligibility maneuver, which is unacceptable as well. Yablo’s theory, Eklund’s theory, and my theory are naturally internalizable. However, both Yablo’s theory and Eklund’s theory are unacceptable because they employ the inconsistent concept of truth. Furthermore, both theories are self-refuting as well. One can construct versions of them that avoid these problems, but they are not naturally internalizable. The confusion-based theory I endorse is naturally internalizable, it is descriptively correct, it does not employ an inconsistent concept, and it poses no self-refutation problems or revenge paradoxes; it is the only one of the theories I have considered to have these properties.

APPENDIX E

THEORIES OF INCONSISTENT CONCEPTS

E.1 INTRODUCTION

In Chapters Four, Five, and Six, I presented a theory of inconsistent concepts, and in Chapter Seven, I used this theory to arrive at a theory of truth on which truth is an inconsistent concept. In this paper, I compare and contrast the theory of inconsistent concepts I develop with several others in an effort to justify the claim that the theory I offer is better than its competitors. In section two, I discuss inconsistent concepts, I present an example of an inconsistent concept that is due to Anil Gupta, and I propose four conditions on acceptable theories of inconsistent concepts. In section three, I discuss the prospects for treating inconsistent concepts as if they are context dependent. In section four, I present an account of indirect context dependence and evaluate its efficacy as a theory of inconsistent concepts. In section five, I address a theory that treats linguistic expressions for inconsistent concepts as if they are ambiguous. In section six, I evaluate two theories that employ supervaluation semantics for inconsistent concepts. The first theory is Hartry Field's; it assigns semantic values to sentences that express the concept in question by treating it as indeterminate and considering all the various ways of making it determinate. Matti Eklund's theory is the second, and it assigns semantic values to sentences that express the concept in question by considering whether the assignment satisfies a weighted majority of the meaning-constitutive principles for the concept. In section seven, I discuss Yablo's theory of inconsistent concepts, which is based on a theory of circular definitions. In section eight, I evaluate dialetheism (which is the view that some sentences are both true and false) as a theory of inconsistent concepts. In section nine, I consider a suggestion of Gupta's for how to interpret inconsistent concepts, which involves the notion of a frame of interpretation. Finally, in section ten, I present an outline of the theory I accept. I argue that it is the only one to satisfy the four constraints on theories of inconsistent concepts (which can be found in section E.2.2).

E.2 Inconsistent Concepts

I discuss the distinction between consistent and inconsistent concepts at length in Chapter Four; in this section, I present a brief overview. On my view, concepts have constitutive principles. These principles are rules that specify how the concept is to be employed. Simply by employing the concept, a person is obligated to follow these rules (of course, it is not the case that everyone actually obeys the constitutive principles for a given concept—some disobey out of ignorance, others do so on purpose). Some concepts have constitutive rules that are compatible in the sense that one can follow all the rules in every situation in which the concept is employed. Other concepts have constitutive rules that are incompatible in the sense that it is impossible to follow all the rules for employing the concept because, in some situations, the rules demand that the employer of the concept use it in incompatible ways simultaneously.

In Chapter Four, I present several distinctions that are relevant to inconsistent concepts, and I provide a number of examples, including:

(1a) 'rable' applies to x if x is a table.

(1b) 'rable' disapplies to x if x is red.

Rable is an inconsistent concept. Someone who possesses **rable** might run into difficulty employing it because it both applies and disapplies to red tables. When confronted with a red table, an employer of **rable** will be unable to satisfy the demands it places on her. Of course, someone could employ **rable** without trouble as long as she avoids red tables. However, even if an employer of **rable** never encounters a red table, the concept still poses a problem for her because inconsistent concepts pose a normative problem for their employers. Someone who chooses to employ **rable** *should* apply it to tables and *should* disapply it to red things. These are conceptual norms to which the employer has decided to bind herself. Thus, an employer of **rable** has committed herself to obeying incompatible rules even if she never encounters a red table.

In the first subsection, I present a more elaborate example of an inconsistent concept, which serves as my example throughout this paper. In the second subsection, I propose four conditions on theories of inconsistent concepts.

E.2.1 The Higherians

The following is an example of an inconsistent concept that is inspired by an example in Gupta's provocative paper "Meaning and Misconceptions."[1] Consider a community of people who speak a language that is similar to English except that in their language, the rules for using the

[1] Gupta (1999). See Allen (forthcoming) for discussion.

expression ‘x is up above y’ (where ‘x’ and ‘y’ are replaced by singular terms) are different. I call the members of this community *Higherians*. Two equally important features of the Higherians’ ‘up above’ talk are that they can perceptually distinguish situations in which one object is up above another (these situations are similar to the ones in which an English speaker would say that one object is up above another), and that they can determine when the ray connecting two objects is parallel to a particular ray that is designated as “Standard Up” (where Standard Up is orthogonal to a tangent plane for the surface of the planet on which the Higherians live). An assertion of ‘A is up above B’ is warranted if and only if either A and B are constituents of one of the perceptually distinguishable situations, or the ray connecting A and B is parallel to Standard Up and A is farther from the surface than B. An assertion of ‘A is not up above B’ is warranted if and only if either A and B are not in the proper perceptually distinguishable relation to one another, or it is not the case that both the ray connecting A and B is parallel to Standard Up and A is further from the surface than B.

Assume that ‘up above’ is defined only for perceivable objects and only for objects within the national borders of the Higherians’ country. When a Higherian can perceive two objects at the same time, then that person can perceive whether they are in the right perceptually distinguishable relation to one another. In addition, every Higherian can determine the ray that connects any two perceivable objects and can determine whether any two rays are parallel. Thus, if a Higherian can perceive object A and he can perceive object B (not necessarily simultaneously), then he can determine whether the ray that connects them is parallel to Standard Up. Assume that the Higherians do not know that their concept is inconsistent because when they can perceive two objects at the same time, they employ the perceptual criterion (i.e., they determine whether the objects are in the proper perceptually distinguishable relation) and when they cannot, they employ the conceptual criterion (i.e., they determine whether the ray connecting the two objects is parallel to Standard Up and they determine which object is closer to the surface of the planet on which they live). Assume also that whether one object is up above another does not depend on any of the Higherians taking them to be in this relation and that the notion of warrant is not relative to anyone's epistemic situation. Finally, assume that there is no difference between the Higherians' idiolects and their common language, that there is no conversational implicature associated with statements containing 'up above', and that the conventions governing 'up above' are common knowledge (i.e., there is no division of linguistic labor for this expression).

If the Higherians live on the surface of a spherical planet then 'up above' is inconsistent. If A and B are two objects that are located some distance from where Standard Up intersects the surface of their sphere and are in the right perceptually distinguishable relation then both 'A is up above B' and 'A is not up above B' will count as warranted because they are in the right perceptually distinguishable relation but the ray connecting them is not parallel to Standard Up. However, if the Higherians' country is confined to one flat surface of a rectangular solid, then 'up above' is consistent because it is defined only within their national borders. I call concepts like **rable** that are inconsistent by definition *intrinsically inconsistent*, and I call concepts like **up above** (in the case where the Higherians live on the surface of a sphere) *empirically inconsistent*. From here on, I assume that the Higherians do indeed live on the surface of a sphere.

Gupta argues that neither conceptual role theories of meaning (i.e., those that explain the meaning of a sentence in terms of rules governing its proper use) nor representational theories of meaning (i.e., those that explain meaning in terms of relations between linguistic and nonlinguistic entities) can give a proper account of the meaning of sentences containing 'up above' locutions. According to Gupta, a proper account should provide both a way to distinguish between true (or warranted)[2] and false (or unwarranted) sentences that contain 'up above' and a way to distinguish between valid and invalid arguments that contain such sentences.

Gupta gives two examples of how people in the community use 'up above' to successfully guide their actions. The first is the *lamp example* in which two roommates disagree about which lamp in their kitchen is broken. Tim asserts 'the lamp up above the stove is broken'. Helen denies this claim. It turns out that although the ray connecting the broken lamp and the stove is not parallel to Standard Up, the lamp and the stove are in the proper perceptually distinguishable relation to one another. Gupta claims that in this case, it is legitimate to say that Tim's assertion is true (or warranted). He provides another example (the *Vishnu example*) in which a person's assertion involving 'up above' is true (or warranted) by virtue of the conceptual criterion instead of the perceptual one.[3]

The problem for conceptual role semantics (on Gupta's account) is that it must treat assertions of sentences whose components are governed by incompatible rules as both warranted and unwarranted. Obviously, any assertion that one object is up above another will be both warranted and unwarranted if the two objects satisfy one criterion for 'up above' but fail to satisfy the other. For example, if the ray connecting two objects is not parallel to Standard Up, but they are in the right perceptually distinguishable relation to one another, then an assertion that one is up above the other will be both warranted and unwarranted. Hence, conceptual role semantics cannot account for the fact that Tim's assertion in the lamp example is true (or warranted). It seems possible (and indeed likely) that some of our own expressions are

[2] Gupta uses these two options and I will follow him. The point is to accommodate assertibility condition theories of meaning.
[3] Gupta (1999: 21).

analogous to ‘up above’; yet we seem to be able to use them without judging that any assertion involving them is both warranted and unwarranted.[4]

According to Gupta, representational theories of meaning do not fare any better because they cannot specify which relation ‘up above’ represents. Instead of arguing for this directly, Gupta considers several potential relations and finds problems with each one. One proposal is that ‘up above’ represents the relation of being parallel to Standard Up. Another is that it represents the relation that obtains between two objects when the ray connecting them passes through the center of the earth and the second object is closer to the center of the earth than the first. Gupta finds three problems with these proposals. First, they privilege one criterion over the other in all situations. Consequently, they provide the wrong assessments of many assertions involving ‘up above’. Moreover, there seems to be no good reason to prefer one of these interpretations to the other. Finally, someone who endorses one of these interpretations will have a difficult time explaining why some assertions involving ‘up above’ are good guides to action while their negations are not.[5]

E.2.2 CONDITIONS ON ACCEPTABLE THEORIES

In this subsection, I present four conditions that any acceptable theory of inconsistent concepts should meet. They are: (i) the theory should imply that the concepts in question are genuinely inconsistent (e.g., it should not reinterpret the concepts so that they have some other semantic features), (ii) the theory should be inferentially charitable (i.e., the theory should not imply that those who employ inconsistent concepts are poor reasoners), (iii) the theory should permit one to distinguish between concept possession and concept employment, and (iv) the theory should

[4] Gupta (1999: 19-21).

[5] Gupta (1999: 22-26)

apply to both intrinsically inconsistent concepts and empirically inconsistent concepts. I discuss the conditions in order.

The first condition is that a theory of inconsistent concepts should be a theory of *inconsistent concepts*—it should not imply that there are no such things or that what we take to be an inconsistent concept is really some type of consistent concept. I argue in Chapters Four and Six that the strategy of reinterpreting a linguistic practice so that what seems to be the employment of an inconsistent concept might work in certain cases, but it fails as a general policy for handling inconsistent concepts. I have no doubt that, given the brief description of the Higherians' linguistic practice, one could plausibly interpret their concept **up above** as some sort of consistent concept (e.g., as context dependent, as vague, as circular, as intensional). However, I also have no doubt that one could present a new example that is very much like the example given in the previous subsection except that **up above** cannot be plausibly interpreted as the sort of consistent concept in question. Thus, given that one can construct genuinely inconsistent concepts, we need a theory of such things.

A theory of inconsistent concepts should include at least: (i) an account of the distinction between consistent and inconsistent concepts, (ii) conditions on the logic that should be used to classify arguments that display inconsistent concepts as valid or invalid, (iii) conditions on a pragmatic theory that applies to speech acts involving inconsistent concepts, (iv) conditions on the semantic theory that applies to sentences that express inconsistent concepts, and (v) a policy for handling inconsistent concepts (i.e., a strategy to follow for those who discover that one of their concepts is inconsistent). Some of the theories of inconsistent concepts I discuss below do not include all five parts, but I do not fault them on these grounds. However, if it seems that a particular theory cannot be amended to include one of these parts, then that is a serious problem.

The second condition is that a theory of inconsistent concepts should be charitable. In particular, a theory of inconsistent concepts is unacceptable if it implies that those who employ inconsistent concepts are irrational. There are plenty of types of rationality and I do not discuss them all here. One type of rationality on which I want to focus is inferential rationality. A theory of inconsistent concepts has implications for the inferential rationality of those who employ inconsistent concepts. Given that an account of inconsistent concepts should include a logic for inconsistent concepts, when one adopts a certain theory of inconsistent concepts, one decides how to treat the reasoning practice of people who employ inconsistent concepts. Thus, when one adopts a theory of inconsistent concepts, one undertakes a commitment to evaluate arguments in which such concepts are expressed according to a certain standard and to treat people who employ such concepts as if they should reason according to that standard.[6]

Although I do not claim to have an exhaustive list, some of the aspects of inferential rationality include being able to determine when arguments are valid or invalid, being able to determine when inductive arguments are strong or weak, being able to weigh evidence for and against a claim, having the capacity and motivation to follow inference rules in one's reasoning, and having the capacity and the motivation to alter one's beliefs effectively in light of conflicting evidence. One can employ an inconsistent concept and still be inferentially rational in all these ways. A theory of inconsistent concepts should respect this fact.

In particular, a theory of inconsistent concepts should imply that a person who employs an inconsistent concept is: (i) capable of following the formal inference rules he accepts, (ii) capable of following the formal inference rules of the logic used to evaluate his arguments, (iii) motivated to follow the formal inference rules of the logic used to evaluate his arguments, (iv) capable of following the material inference rules he accepts (i.e., capable of following his

[6] See Camp (2002) and Chapter Five for discussion.

accepted strategies for weighing evidence), (v) capable of following the material inference rules of the semantic theory used to interpret his utterances and beliefs, and (iv) motivated to follow the material inference of the semantic theory used to interpret his utterances and beliefs.[7]

Two features of inconsistent concepts make the inferential rationality condition on theories of inconsistent concepts especially urgent. First, the potential for paradoxical reasoning accompanies the employment of an inconsistent concept. For example, let *R* be the name of a red table. R is a table; hence, R is a rable. R is red; hence, it is not the case that R is a rable. Thus, R is a rable and it is not case that R is a rable. We have arrived at a contradiction via intuitively plausible steps from intuitively plausible assumptions. Consider another example. Assume for reductio that some red tables exist. Let R be the name of a red table. The reasoning above shows that R is a rable and R is not a rable. Contradiction. Therefore, no red tables exist. We have proven an obviously false sentence via intuitively plausible steps from intuitively plausible assumptions. If one accepts classical logic and treats 'rable' as univocal and invariant, then one will have a hard time avoiding these unacceptable conclusions. Hence, there is considerable pressure to endorse non-classical logics for evaluating arguments that involve inconsistent concepts. Second, a person can possess and employ an inconsistent concept without knowing that it is inconsistent. Indeed, anyone who discovers that one of his or her concepts is inconsistent should cease employing it. Thus, a theory of inconsistent concepts will be used primarily to interpret people who are using an inconsistent concept without knowing that it is inconsistent. Given that most employers of inconsistent concepts are ignorant of their inconsistency and that many theories of inconsistent concepts include non-standard logics for

[7] In Chapter Five, I argue that the inferential rationality condition implies that a theory of confused concepts should imply that sentences that express confused concepts do not have truth values. An analogous argument shows that sentences that express inconsistent concepts do not have truth values.

inconsistent concepts, the potential for treating those who employ inconsistent concepts as inferentially irrational is high.

It might seem impossible for a concept to be inconsistent without the employers of that concept knowing that it is inconsistent, but the fact that concepts can be empirically inconsistent should dispel this impression. The rules for the employment of a concept often incorporate features of the environment in which it is used; if the employers of a concept are ignorant or mistaken about some features of their environment, then the concept in question can be inconsistent without their knowledge. No amount of "reflection on their concepts" will inform them that their concept is inconsistent; they have to go out into the world and discover empirical facts to discover the conceptual inconsistency. Consider the history of human inquiry—we (humans) discover false empirical beliefs alarmingly often. Given the extent of our ignorance and error, there is a good chance that many, perhaps most, of our concepts are empirically inconsistent. That sobering thought should lend urgency to the task of constructing an adequate theory of inconsistent concepts.

The third condition is that a theory of inconsistent concepts should permit one to distinguish between concept possession and concept employment. This distinction is not important for acceptable concepts—any acceptable concept I possess is a concept I employ. Here, employing a concept does not mean actively applying it or disapplying it. Rather, employing a concept is being ready and willing to apply or disapply it should an occasion arise. However, not all concepts are acceptable; some are defective. Indeed, an inconsistent concept is a particular type of defective concept. When it comes to defective concepts, the distinction between employment and possession is essential. *Ceteris paribus*, one should stop employing a concept one takes to be defective. Of course, when one decides to stop employing a particular concept, one still possesses it. Given that there is a valid distinction between defective and acceptable concepts and that one ought to stop employing a concept one believes to be defective (despite the fact that one still possesses it), there is an important distinction between concept possession and concept employment. A theory of defective concepts in general, and a theory of inconsistent concepts in particular should respect this distinction.

Some remarks of Dummett's on concepts will illustrate this distinction. Dummett's account of concepts employs the distinction between the circumstances of application and the consequences of application.

> The distinction is thus meant as no more than a rough and ready one, whose application, in a given case, will depend in part on how we choose to slice things up. It remains, nevertheless, a distinction of great importance, which is crucial to many forms of linguistic change, of the kind we should characterize as involving the rejection or revision of concepts. Such change is motivated by a desire to attain or preserve a harmony between the two aspects of an expression's meaning. A simple case would be that of a pejorative term, e.g., 'Boche'. The condition for applying the term to someone is that he is of German nationality; the consequences of its application are that he is barbarous and more prone to cruelty than other Europeans. … Someone who rejects the word does so because he does not want to permit a transition from the grounds for applying the term to the consequences of doing so, (Dummett 1973: 454).

One important feature of Dummett's model is that it permits a characterization of defective concepts like **Boche**. I possess this concept; however, I do not employ it. I do not employ this concept because I disagree with it in some sense. In particular, I reject the inference from its conditions of application to its consequences of application. Because I disagree with it in this sense, I do not formulate either positive or negative judgments with it—I do not apply it and I do not disapply it. I reject both the claim that some person is Boche and the claim that he is not Boche. However, I certainly possess it. I can attribute it to others and I understand claims made with it.[8]

[8] See also Brandom (1994: 116-130).

Without the ability to distinguish between concept employment and concept possession, it is impossible to give a plausible account of how one person can attribute an inconsistent concept to another without falling into inconsistency herself. Furthermore, if a theory of inconsistent concepts appeals to inconsistent concepts, then one cannot accept it without employing the inconsistent concepts in question. A theory of inconsistent concepts is acceptable only if it implies both that a person can possess and attribute an inconsistent concept without employing it, and that one can accept a theory of an inconsistent concept X without employing X.

One prominent account of inconsistent concepts has difficulty distinguishing between concept possession and concept employment. This account combines the view that some sentences are meaning-constitutive with the view that the meaning of a linguistic expression is the concept it expresses. A sentence is *meaning-constitutive* for a linguistic expression that occurs in it if and only if the sentence partially defines the linguistic expression in question. The account of inconsistent concepts I have in mind implies that a concept is inconsistent if and only if the set of meaning-constitutive sentences for the linguistic expression that expresses the concept is inconsistent.[9]

Before presenting the problem for this account of inconsistent concepts, I want to discuss an ambiguity in the account of meaning-constitutivity. Assume that 'w' is the name of a word, 'p' is the name of a sentence, and 'm' is the name of a meaning such that w occurs in p, and w means m. If one claims that p is meaning-constitutive for w, then one could mean at least two different things. First, one could mean that for a person's use of w to mean m, that person must believe the proposition expressed by p. This interpretation permits p to be both meaning-

[9] See Peacocke (1993, 2000), Boghossian (1996, 1997, 2000), Horwich (1998), and Hale and Wright (2000) for examples of meaning-constitutive theories. See Chihara (1979, 1983), Priest (1987), Gupta (1999), and Eklund (2002) for examples of meaning-constitutive theories of inconsistent concepts.

constitutive and false. Second, one could mean that for w to mean m, p must be true. This interpretation permits the person in question to disbelieve the proposition expressed by p. I refer to the first type of meaning-constitutivity as *cognitive* and the second as *aletheic*.[10]

These two types of meaning-constitutivity correspond to two accounts of inconsistent concepts. On the first, the cognitive meaning-constitutive sentences associated with the concept are inconsistent; for the second, the aletheic meaning-constitutive sentences associated with the concept are inconsistent. According to the second account, no one could ever express an inconsistent concept because for one's use of some expression to express an inconsistent concept, the meaning-constitutive sentences associated with that concept must be inconsistent and true. However, no set of true sentences is inconsistent.[11] Hence, if meaning-constitutive accounts of inconsistent concepts are to be helpful or explanatory at all, then they must be taken in the cognitive sense.

One can give a similar account of inconsistent concepts based on inference rules instead of sentences. On this account, a concept is inconsistent if and only if the set of meaning-constitutive inference rules associated with it is inconsistent (i.e., it is not the case that they can all be valid). A similar ambiguity haunts this conception as well. A set of inferences could be meaning-constitutive for some concept in the sense that that one has to endorse them to possess that concept, or they could be meaning-constitutive in the sense that they must be valid. As with the sentence-based accounts, only one of the inference-based accounts is plausible; I refer to it as *cognitive* as well.

According to the cognitive meaning-constitutive account, possession of an inconsistent concept is explained in terms of holding all the beliefs that belong to an inconsistent set or

[10] Eklund (2002) contains some gestures toward this distinction.

[11] I am ignoring dialetheism for the moment.

endorsing all the inference rules that belong to an incompatible set. I use the term ‘principles’ as a generic term for beliefs and inference rules. Thus, on both cognitive meaning-constitutive accounts, if a person possesses an inconsistent concept, then he accepts all its meaning-constitutive principles.

Consider how a cognitive meaning-constitutive account of inconsistent concepts applies to the Higherians. Assume that U is the set of inconsistent meaning-constitutive principles for **up above**. If the cognitive meaning-constitutive account is correct, then it is impossible for someone to possess **up above** without accepting all the members of U. Thus, anyone who employs **up above** accepts all the members of U. However, employing a concept is not the only thing one can do when one possesses it. Indeed, one can attribute it to someone else. In order to attribute an inconsistent concept, one must possess it. Thus, anyone who attributes **up above** to someone else accepts all the members of U. In particular, an interpreter of the Higherians must accept all the members of U in order to attribute **up above** to them. This cannot be right. One should not have to endorse an inconsistent set of principles to attribute an inconsistent concept to someone else. Furthermore, if someday the Higherians are lucky enough to realize that their concept is inconsistent and they decide to stop using it, then according to the cognitive meaning-constitutive account, they would no longer possess it. Again, this is surely wrong. Even if they change their beliefs and decide to no longer *employ* **up above**, they still *possess* it. Therefore, a person should be able to attribute an inconsistent concept without accepting all the members of an inconsistent set of principles. The cognitive meaning-constitutive account of inconsistent concepts does not respect the distinction between concept possession and concept employment. Without this distinction, I see no hope for a plausible account of inconsistent concepts.

A consequence of this condition on theories of inconsistent concepts is that sentences that express inconsistent concepts do not have truth values. Let T be a theory of an inconsistent concept X, and let T imply that a sentence p, which expresses X, has a truth-value. Let p be 'α is X'. Assume that a person S who possesses X accepts T. If T implies that p is true, then S accepts that p is true. If S accepts that p is true, then S accepts that α is X. Hence, if T implies that p is true and S accepts T, then S employs X. On the other hand, if T implies that p is false, then S accepts that p is false. If S accepts that p is false, then S accepts that α is not X. Hence, if T implies that p is false and S accepts T, then S employs X. Therefore, if T implies that p is either true or false and S accepts T, then S employs X. Any theory of inconsistent concepts that implies that sentences that express inconsistent concepts have truth-values is unacceptable to someone who refuses to employ inconsistent concepts. Consequently, any acceptable theory of inconsistent concepts implies that sentences that express inconsistent concepts are neither true nor false.[12]

The fourth condition on acceptable theories of inconsistent concepts is that they should apply to both essentially inconsistent concepts and empirically inconsistent concepts. The example given in section E.2.1 of the Higherians and their concept **up above** shows that an account of concepts that are inconsistent by definition is not enough. One must be able to explain concepts that turn out to be inconsistent because of the environment in which they are used.

[12] Camp argues that this condition follows from the inferential rationality condition as well; I discuss it in Chapter Five.

E.3 Context Dependence

In this section, I consider an account of inconsistent concepts that implies that they are context dependent. Gupta considers one such proposal on which ‘up above’ is an implicit indexical. The suggestion is that the interpretation of ‘up above’ talk is relative to the context in which it is produced. It seems that the best way to carry out this suggestion is to alter the criteria by which the sentences that contain ‘up above’ are evaluated. The conceptual criterion should be based on the direction of Standard Up as determined by the location of the discussion (the ray from the center of the Earth through the object closer to it). Gupta follows Kaplan in thinking of the meaning of sentences that contain ‘up above’ as a function from context to content.[13] One can then evaluate the content of sentences that contain ‘up above’. According to this suggestion, someone’s use of ‘up above’ will be evaluated by either the perceptual criterion or the conceptual criterion if her utterance takes place in the vicinity of the objects to which she refers (the two criteria should overlap) and by the conceptual criterion if her utterance takes place elsewhere.

Gupta points out that the indexical semantics will provide the wrong verdict for claims made away from their subject matter. The indexical semantics often attributes the wrong truth-value to sentences containing ‘up above’ that are uttered some distance from the objects to which it refers. This fact renders the practices of reassertion and appeal to authority useless in discussions involving ‘up above’. Gupta also claims that propositional attitude attributions that contain ‘up above’ are not incomplete in the sense that they require some contextual element to

[13] Kaplan (1989).

fix their meaning. Furthermore, the indexical theory implies that the contents of the Higherians' beliefs about the relative positions of objects change as they move around the world.[14]

The second and third problems seem to me to be less pressing than the first one because propositional attitude ascriptions that contain indexicals pose a general problem that is not specific to the Higherians and 'up above'. However, the first problem is serious and seems to me to sink the indexical semantics. Nevertheless, there are other models of context dependent terms that do not make the semantic features of a word depend on the location of its utterance. For example, one could say that the content of 'up above' is determined by the common knowledge of those participating in the conversation in which it is uttered.[15] The common knowledge in the conversation would determine whether the sentence containing 'up above' should be evaluated by the conceptual criterion or the perceptual one. If one posited a conceptual criterion for every location then such a theory might do a better job than the indexical ones.

A significant problem would remain and it seems to me to be the most compelling reason to reject context dependent accounts of inconsistent concepts. The problem is that linguistic expressions have their semantic features because of the way they are used. If a linguistic expression is context dependent, then it has been used in a way that renders it context dependent. One might be able to make a case for the claim that, as described, the Higherians use 'up above' in a way that renders it context dependent. However, one could present a new example in which the members of the linguistic practice do not use the expression in question in this way. Therefore, the context dependence theory does not satisfy the first condition on theories of inconsistent concepts; i.e., it reinterprets inconsistent concepts as consistent concepts that are context dependent.

[14] Gupta (1999: 24-26).

[15] See Stalnaker (1973).

I am *not* claiming that everyone who uses a context dependent expression must know that it is context dependent. Indeed, if one asks English speakers whether 'tall' is context dependent, then many will say "no." Of course, even the people who deny that 'tall' is context dependent will treat it as context dependent, and once one gives them the appropriate examples, chances are that they will accept that 'tall' is context dependent.

My point is that a theory of inconsistent concepts on which inconsistent concepts are context dependent is unacceptable as a general theory because it implies that some linguistic expressions display context dependence even though the members of the linguistic practices to which these expressions belong do not use them in a way that renders them context dependent. I presented the Higherians as a people who use an inconsistent concept and I resist any attempt to reinterpret 'up above' to eliminate the inconsistency.

E.4 Indirect Context Dependence

One natural change to make to the context dependence theory is to allow the evaluation of sentences containing 'up above' to be relative to the context that contains the objects referred to by the singular terms in the sentence, rather than relative to the context of utterance. I refer to this as the *indirect context dependence theory*. How exactly would such a theory work? Let us assume that Tim utters 'A is up above B' where 'A' and 'B' are names for objects. Assume as well that the conversation takes place away from A and B. What criteria should be used to evaluate Tim's sentence? Here is a suggestion. If a normal observer in standard conditions near B could perceive both A and B simultaneously, then the perceptual criterion should be used, and it should be based on this ideal observer. If A and B are not simultaneously perceivable by such an observer, then the conceptual criterion should be used.

This indirect context dependence theory surely works better than the one based on treating 'up above' as context dependent. However, one problem for the former is simply the lack of any clear theory to govern it. For indexicals, we have several explicit theories from which to choose. It is unclear whether we even have any terms that behave like 'up above' on the indirect context dependence theory. I like to think of the difference between the behavior described by the two semantic theories as analogous to the difference between the behavior of an indexical and the behavior of words like 'illegal'. If I say 'X occurred here' then the truth-value of my claim should be interpreted relative to my position. However, if I say 'X is illegal' then the truth-value of my claim should be determined not relative to my position when I uttered it, but rather relative to the place in which X was performed. Tim's utterance of 'the action of the intruder who broke into my home in 1979 was illegal' does not change based on the legal system in which he utters it. Instead, it depends on the legal system that applies to the place where the action occurred. This analogy does little to reply to the objection at hand. It seems that 'up above' on the indirect context dependence theory and 'illegal' as it is used in English share some common features. However, there is still no general theory to govern this behavior.

Another problem is that the modified contextual semantics might work well for assigning truth-values to the Higherians' sentences that contain 'up above', but it hardly constitutes a solution to the problem of providing a theory of inconsistent concepts. I do not see any clear way of applying it to other examples of inconsistent concepts (e.g., **rable** and the other examples in Chapter Four).

E.5 Ambiguity

In this section, I want to consider the theory of inconsistent concepts on which linguistic expressions that express inconsistent concepts are ambiguous. There are at least two distinct theories of this sort. The first theory distinguishes between a number of different meanings of the expression in question and posits a principle for assigning one of these meanings to the linguistic expression (i.e., a disambiguation principle). Any number of different logics and semantic theories are compatible with this theory; presumably, the theorist would use whichever logic and whichever semantic theory is appropriate for the discourse in question. The second theory implies that the expression in question is ambiguous, but it does not disambiguate. Instead, it classifies an argument containing the expression in question as valid if and only if it would be valid no matter how the expression is disambiguated. In this section, I consider only the first type of theory because the second is equivalent to a version of the dialetheic theory of inconsistent concepts, which is the topic of section eight.[16] I call the first type a *disambiguation theory*.

I assume that in the case of the Higherians, the most natural way to construct a disambiguation theory would be to posit one meaning for the perceptual criterion and another for the conceptual one. I use the terms ‘perceptually up above’ and ‘conceptually up above’ as expressions with these meanings, respectively. Any disambiguation theory will have to provide an account of which meaning should be assigned to ‘up above’ for each assertion in which it occurs. I call this a *disambiguation principle*. I suggest that we use the asserter’s dispositions to justify the assertion if challenged. For example, assume that Tim asserts ‘A is up above B’. If Tim were challenged to justify this assertion, then he would most likely either say that he

[16] See Lewis (1982), Priest (1995), and Allen (forthcoming).

perceives that A and B stand in this relation, or say that the ray connecting A and B is parallel to Standard Up. In the former case, ‘up above’ had the same meaning as ‘perceptually up above’ and in the latter it had the same meaning as ‘conceptually up above’. If a person’s assertion of a sentence containing ‘up above’ deserves the perceptual reading, then it is true just in case the perceptual analog is true; if a person’s assertion of a sentence containing ‘up above’ deserves the conceptual reading, then it is true just in case the conceptual analog is true.[17]

Presumably, a disambiguation theory for ‘up above’ would respect most of the intuitive assessments Gupta employs. In particular, it provides the right truth-values for the lamp example and the Vishnu example. Thus, it does better than any of the other options Gupta considers. However, there is a major problem with it. It fails to draw an acceptable distinction between valid and invalid inferences that contain ‘up above’ sentences. Because the Higherians do not know that ‘up above’ is ambiguous, they will often present and endorse arguments that contain sentences to which a disambiguation theory gives different readings. For example, Tim might endorse the following argument:

A is up above B.

B is up above C.

For all x, y, and z, if x is up above y and y is up above z, then x is up above y.

∴ A is up above C.

If the first premise should receive the perceptual up above interpretation and the second the conceptual up above interpretation, then the disambiguation theory entails that this argument contains an equivocation. Thus, a disambiguation theory implies that many of the Higherians’ arguments contain equivocations. Therefore, it is inferentially uncharitable. In particular, it

[17] This account is for illustration purposes only. The points I make about ambiguity semantics hold regardless of one’s choice of meanings and how to disambiguate.

implies that the Higherians are not capable of following the inference rules they endorse. Hence, the disambiguation theory does not satisfy the second condition on theories of inconsistent concepts.[18]

E.6 SUPERVALUATION

In this section, I present two theories of inconsistent concepts that employ supervaluation. Supervaluation is a logical technique for assigning semantic values to sentences that display some sort of indeterminacy. In the first subsection, I present the theory Field uses in his early account of referential indeterminacy. In the second subsection, I present Eklund's theory, which is designed for inconsistent concepts.

E.6.1 FIELD'S THEORY

Hartry Field began his career endorsing a version of the correspondence theory of truth. This theory fits neatly with representational account of meaning, on which each singular term is linked via a relation (often called *reference*) between it and an object in the world and each predicate is linked via a relation (often called *denotation*) between it and a set of objects in the world. Field then raised a problem for this account: the history of science often displays a phenomenon Field calls *referential indeterminacy*. (It is essential to distinguish this phenomenon from *indeterminacy* (*or inscrutability*) *of reference*, which is a doctrine made famous by Quine and Davidson to the effect that one can rearrange all the references of our

[18] This criticism is similar to the one Camp presents of ambiguity interpretations of confusion; see Chapter Five.

words without disturbing the truth values of the sentences in which they occur.[19]) In Chapter Five, I discussed Field's example of referential indeterminacy, which involves the term 'mass' as it was used in Newtonian physics. Recall that the problem is to specify the link between the term 'mass' and the nonlinguistic elements of the world that explains how it functions.

Field's solution is to use a supervaluation semantics for sentences involving 'mass' in Newtonian physics. That is, he treats 'mass' as if it *partially denotes* both relativistic mass and proper mass, but it does not *fully denote* either. To determine whether a sentence with 'mass' in it is true, one must evaluate two other sentences, one with 'relativistic mass' in place of 'mass' and one with 'proper mass' in place of 'mass'. If both of these sentences are true, then the original is true too. If both are false, then the original is false as well. However, if one is true and one is false, then the original sentence is neither true nor false. This method is called *supervaluation*; it is a popular strategy for dealing with a number of different types of defective discourse (e.g., reference failure[20], presupposition failure[21], vagueness[22], and the semantic paradoxes[23]).

A supervaluation semantics for 'up above' discourse would treat 'up above' as partially representing multiple relations (just as 'mass' partially denotes several properties). The most natural choice is to use the perceptual up above relation and the conceptual up above relation. On this semantics, to determine whether a sentence with 'up above' is true, one must evaluate two other sentences, one with 'perceptually up above' in place of 'up above' and one with 'conceptually up above' in place of 'up above'. If both sentences turn out true, then the original

[19] See Quine (1960) and Davidson (1973).
[20] van Fraassen (1967).
[21] van Fraassen (1968).
[22] Fine (1974).
[23] van Fraassen (1968, 1970), Kripke (1975), McGee (1991).

sentence is true; if both turn out false, then the original sentence is false. If one is false and one is true, then the original sentence is neither true nor false.

It seems that a supervaluation semantics will do a poor job of conforming to the Higherians' intuitions. The only sentences that turn out true will be those pertaining to objects connected by the Standard Up ray. That does not seem right. In particular, the supervaluation semantics does not provide the right verdict on either the lamp example or the Vishnu example. For a sentence containing 'up above' to count as false, it must fail both criteria for 'up above'. Presumably, many of the sentences that are intuitively false will count as false on the supervaluation approach. Virtually all of the sentences containing 'up above' that are intuitively true will be neither true nor false. Given that this theory implies that some of the sentences containing 'up above' have truth-values, it fails to satisfy the third condition on theories of inconsistent concepts.[24]

E.6.2 EKLUND'S THEORY

Eklund provides a theory of truth and a theory of vague concepts on which truth and vague concepts are inconsistent. He argues that by virtue of our semantic competence, we accept the premises and the inference rules that lead to the liar paradox and the sorites paradox. Eklund's theory is a cognitive meaning-constitutive theory. That is, an inconsistent concept is one for which the associated set of principles that one must accept in order to possess the concept are inconsistent. He resists the temptation to think of meaning-constitutive sentences as true or unrevisable. To clarify this claim, he introduces the notions of competence dispositions and

[24] There are several different validity criteria for supervaluation semantics, and I do not have the space to discuss them here, but it seems to me that they will pose problems that are similar to the problem with the disambiguation theory. See Keefe (2000), Michael (2002), and Kremer and Kremer (2003).

culprits. One's *competence dispositions* are belief-forming dispositions that one has by virtue of one's semantic competence. A *culprit* is the false premise or invalid inference used in the derivation of the contradiction in a paradox. One can say that one's competence dispositions lead one to accept the culprit of the paradox because the set of cognitive meaning-constitutive sentences associated with the concept in question is inconsistent. If a concept displays this phenomenon, then the paradox associated with it is said to *exert pull*. One of Eklund's central theses is that the liar paradox and the sorites paradox exert pull (I call this the *pull exertion thesis*).[25]

Eklund considers the compatibility of this claim with three popular theories of semantic competence: the truth conditional theory, conceptual role semantics, and the Fregean theory.[26] According to the first, to know the meaning of a sentence is to know its truth conditions. For conceptual role semantics, to know the meaning of a sentence is to have the right belief-forming dispositions. Fregean theories imply that the semantic values of sentences satisfy their senses. He argues that all three preclude the pull exertion thesis. The first would require an inconsistent set of truths, the second would require an inconsistent set of valid inference rules, and the third would require semantic values to satisfy an unsatisfiable set of conditions. Eklund also issues a challenge to theories of semantic competence that is similar to Gupta's challenge to theories of meaning. For if the liar paradox and the sorites paradox do exert pull then there must be something wrong with these three accounts of semantic competence.

To remedy this situation, Eklund suggests some changes for conceptual role theories and for Fregean theories. The remedy for the former is to allow competence dispositions for the acceptance of inferences that are not truth preserving and that are defeasible. Fregean theories

[25] Eklund (2002).

[26] Eklund obviously endorses the common view that theories of meaning and theories of semantic competence are closely linked.

need to allow the sense of an expression to be the cognitive meaning-constitutive principles associated with it. The sense of an expression then determines that the semantic value of an expression is whatever comes closest to satisfying these principles.[27]

Eklund's suggestion for a semantics for inconsistent concepts is to define an acceptable assignment of semantic values to expressions of a language, L, as one that makes true a weighted majority of the cognitive meaning-constitutive principles for L. Eklund says very little on how to determine whether an assignment is acceptable and on the weighting function that should be used. In a footnote, Eklund compares his strategy with the traditional supervaluation semantics:

> Talk of "acceptable assignments" is familiar from supervaluationist analyses. But note that the acceptable assignments considered here are quite different from the acceptable assignments—SV-assignments, let us call them—the supervaluationist talks about. For the supervaluationist, the reason there are many SV-assignments for a natural language is that the meanings of some expressions are incomplete: they can be extended without being changed, as Kit Fine puts it (1975, p. 267). The SV-assignments correspond to all the possible completions of meanings of expressions of the language. They are not, as the acceptable assignments discussed here, meant to be as faithful as possible to the meanings expressions actually are endowed with. (As illustrated by, for example, the fact that although supervaluationists normally do not accept bivalence, all the particular SV-assignments are bivalent), (Eklund 2002: 265n33).

As I understand it, the big difference (according to Eklund) between his theory and supervaluation semantics is that, for supervaluation semantics, the semantic values of the expressions in question are determined by considering all sorts of assignments and constructing one on the basis of their shared properties, whereas Eklund's theory determines semantic values by considering a bunch of assignments and picking one (or more) from among them on the basis of whether it satisfies the cognitive meaning-constitutive principles associated with the expressions of the language. On the former, an expression has a certain semantic feature if all the various acceptable assignments imply that it does; on the latter, an expression has a certain

[27] Eklund (2002: 260-266).

semantic feature if the claim that it does is the most compatible with its cognitive meaning-constitutive sentences.

It seems to me that this way of putting the difference is misleading. The real difference is just that most supervaluation semantics do not respect the cognitive meaning-constitutive sentences associated with the expressions of a language. That is, a supervaluation semantics treats all meaning-constitutive sentences as aletheic—i.e., ones that must be true. Eklund places a further constraint on what counts as an acceptable assignment. An acceptable assignment must make a weighted majority of the cognitive meaning-constitutive sentences associated with the expressions of a language come out true. In (still) other words, Eklund's version of supervaluation semantics allows sentences that express penumbral connections to turn out false.[28] Another difference seems to be that Eklund does not take a stand on what to do in the face of multiple incompatible acceptable assignments. He suggests that it might be all right to retain bivalence at the cost of accepting multiple incompatible assignments for some expressions.

Eklund's suggestion does make a nice complement to Field's. Recall that Gupta issued separate challenges to representational theories of meaning and to conceptual role theories. Field's supervaluation semantics is decidedly representational, but Eklund shows how to incorporate what is basically the same idea into a conceptual role approach. Together, Field's and Eklund's semantic theories constitute an important reply to Gupta's challenge. If we assume that Eklund's theory provides basically the same results as a standard supervaluation semantics, then both provide a way of determining truth values and validity for the Higherians' use of 'up above'. We also see that one need not endorse representational semantics to have the supervaluation approach at one's disposal.

[28] Tappenden (1993) makes similar recommendations and Fodor and Lepore (1996) point out the tension between meaning-constitutive sentences and supervaluation semantics. I certainly do not agree with Fodor and Lepore's conclusion or their argument strategy, but the tension they point out is a real one.

How could Eklund's semantics be applied to 'up above' as it is used by the Higherians? I assume that the sentences that express the two criteria would count as cognitive meaning-constitutive. The paradox associated with 'up above' is that for any two objects, A and B, that are some distance from Standard Up and in the proper perceptual relation, one can argue that 'A is up above B' is both true and false. The paradox exerts pull because the Higherians' competence dispositions (i.e., they accept the meaning-constitutive principles associated with 'up above', which are inconsistent) lead them to accept the culprits in the derivation of the contradiction. It seems to me that Eklund's diagnosis provides a nice way of putting the problem.

Eklund's semantic theory dictates that an acceptable assignment of semantic values to the sentences that contain 'up above' will be one that makes true a weighted majority of the cognitive meaning-constitutive sentences associated with 'up above' (I am assuming that 'up above' is the only inconsistent concept in the Higherians' repertoire). Does this constraint help at all? It seems not. We are still in the position of falsifying one criterion or the other and we do not have a good way of determining which one to privilege. Either Eklund's solution reduces to the standard supervaluation solution or it gives us no help in determining the semantic values for expressions of inconsistent concepts.

Another problem with Eklund's theory is that it implies that anyone who possesses an inconsistent concept accepts the meaning-constitutive principles associated with it. Thus, his theory does not respect the distinction between concept possession and concept employment. One consequence is that anyone who endorses Eklund's theory must accept an inconsistent set of principles. Thus, it does not satisfy the third condition on theories of inconsistent concepts.

E.7 Circularity

In this section, I present Yablo's theory of inconsistent concepts. It is based on his theory of circular concepts, which, in turn, owes much to Gupta and Belnap's theory of circular concepts.[29] Yablo presents an account of inconsistent concepts by way of a theory of inconsistent definitions. Let the canonical form of a definition be: $Px =_{\text{def}} \phi(x)$, where $\phi(x)$ is a formula with only 'x' free. He begins by distinguishing between noncircular, positive circular, and negative circular definitions. A *noncircular definition* is one where P does not occur in ϕ, a *positive circular definition* is one where P occurs in φ in such a way that if P's extension increases, ϕ's does as well, and a *negative circular definition* is one where P occurs in ϕ in such a way that if P's extension increases, ϕ's decreases. As for noncircular definitions, we can follow rule (D) to determine their extension on the basis of the definition:

> (D) x satisfies P in world w if and only if x belongs to ϕ's extension in w.

To use (D) one must first determine ϕ's extension in w, which I will assume can be done without trouble for noncircular definitions.

Rule (D) will not work for circular definitions because they require one to first know the extension of P in order to determine the extension of ϕ. One might instead follow rule (E) to determine the extension of P:

> (E) x satisfies P if and only if x is a member of Φ, where Φ is a set that solves the equation $\ulcorner\Phi = \phi_w(\Phi)\urcorner$ and $\phi_w(\Phi)$ is the extension of ϕ in w on the assumption that P has Φ as its extension.

If Φ solves the above equation, then setting P's extension to Φ will make the definition come out true. One problem for using (E) as a rule for determining the extension of circularly defined

[29] Gupta and Belnap (1993).

expressions is that the equation associated with a positive circular definition will usually have multiple solutions and that associated with a negative circular definition need not have any solution at all.[30]

To solve this problem, one can rely on rule (F) for positive circular definitions:

(F) x satisfies P in w if and only if x is a member of Φ, where Φ is a set that constitutes the *least* solution to the equation $\ulcorner \Phi = \phi_w(\Phi) \urcorner$ and $\phi_w(\Phi)$ is the extension of ϕ in w on the assumption that P has Φ as its extension.

The difference between (E) and (F) is that the latter forces the extension of F to be the least solution to the equation associated with F's definition and one can prove that any such equation has a least solution.

For negative circular definitions, one can use:

(G) x satisfies P in w if and only if x is Δ-grounded in w.

Unfortunately, the definition of 'Δ-grounded' is a bit complicated because it incorporates Tarski's definition of satisfaction. The following are the rules that are based on Tarski's definition:

(AT) $\mathbf{s}(x) \in \mathbf{A} \rightarrow T(Ax, \mathbf{s})$.

(AF) $\mathbf{s}(x) \in \mathbf{A} \rightarrow F(Ax, \mathbf{s})$.

(~T) $F(\psi, \mathbf{s}) \rightarrow T({\sim}\psi, \mathbf{s})$.

(~F) $T(\psi, \mathbf{s}) \rightarrow F({\sim}\psi, \mathbf{s})$.

(∧T) $T(\psi, \mathbf{s})$ and $T(\chi, \mathbf{s}) \rightarrow T(\psi \wedge \chi, \mathbf{s})$.

(∧F) $F(\psi, \mathbf{s})$ and $F(\chi, \mathbf{s}) \rightarrow F(\psi \wedge \chi, \mathbf{s})$.

(∀T) $T(\psi, \mathbf{s}')$ for all $\mathbf{s}' \approx_x \mathbf{s} \rightarrow T(\forall x \psi, \mathbf{s})$.

[30] Positive circular definitions are also called inductive definitions, negative circular definitions are also called anti-inductive definitions. See Yablo (1993a: appendix).

($\forall$F) $F(\psi, \mathbf{s}')$ for all $\mathbf{s}' \approx_{x} \mathbf{s} \rightarrow F(\forall x\psi, \mathbf{s})$.

$\ulcorner T(\psi, \mathbf{s})\urcorner$ is synonymous with $\ulcorner$formula ψ is true of the object assigned to its variable by the function, **s**, which assigns objects from the domain to the variables of the language$\urcorner$ and $\ulcorner F(\psi, \mathbf{s})\urcorner$ is synonymous with $\ulcorner$formula ψ is false of the object assigned to its variable by the function, **s**, which assigns objects from the domain to the variables of the language$\urcorner$. Rest assured, these are just the standard clauses from Tarski's definition of satisfaction. I refer to these as the *satisfaction rules*. The rest of the rules are:

(ΔT) $T(\phi, \mathbf{x}) \rightarrow T(P, \mathbf{x})$.

(ΔF) $F(\phi, \mathbf{x}) \rightarrow F(P, \mathbf{x})$.

The first says that one should add x to P's extension if the satisfaction rules prove that ϕ is true of it; the second says that one should not add x to F's extension if the satisfaction rules prove that ϕ is false of it. Now we can define Δ-groundedness:

(Δ-groundedness) An object is Δ-grounded if and only if the satisfaction rules, (ΔT), and (ΔF) prove $T(\phi, \mathbf{x})$.

It turns out that one can rephrase (E) in terms of proof by the satisfaction rules and one can rephrase (F) in terms of the satisfaction rules and (ΔT). Thus, (G) will work for non-circular, positive circular, and negative circular definitions.

There is one more set of definitions that have yet to be dealt with. These definitions are circular but are neither positive nor negative because the definiendum occurs both positively and negatively in the definiens. For these, Yablo suggests a reflection rule that incorporates the reasoning from the groundlessness of the claim that ϕ is true of x to the claim that P is false of x:

(ΔR) $\Theta \rightarrow F(P, x)$, where Θ makes x ungroundable.

Θ is a set of claims about which objects satisfy which formulas and Θ makes x ungroundable if and only if $T(\phi, x)$ is not provable using the satisfaction rules and (ΔT), from the set of F(P, s) such that T (P, s) $\notin \Theta$. Rule (ΔR) allows us to infer that some object is not in P's extension from the fact that $T(\phi, x)$ is not provable from the satisfaction rules and (ΔT) (i.e., it allows us to infer that x is not in P's extension from the fact that $T(\phi, x)$ is groundless).

Finally we get to inconsistent definitions. For Yablo, a definition is *consistent* if and only if (E), (F), and (G) are jointly satisfiable. A definition is *inconsistent* if and only if it is not consistent. As a test of consistency, Yablo defines two sets:

Γ_Δ = {x: the reflective rules prove T(P, x)}

Γ^Δ = {x: the reflective rules do not prove F(P, x)}

Here, the *reflective rules* are the satisfaction rules, (ΔT), and (ΔR). A definition is consistent if and only if $\Gamma_\Delta = \Gamma^\Delta$. Any attempt to follow the semantic rules when dealing with inconsistent definitions is impossible. Yablo draws an analogy with incompatible moral obligations; the difference is that, with inconsistent definitions, an attempt to comply with one obligation creates another that one must defy.[31]

The problems I have with Yablo's theory of inconsistent concepts become apparent as soon as one tries to use it to arrive at a theory of **up above**. The first problem is that, as I have defined it, 'up above' is not circular; I have no idea how to construct a circular definition of it, much less a circular definition that is neither positive nor negative. I agree with Yablo that the concepts defined by the definitions he considers are inconsistent, and his theory might do a good job of handling them. However, there are inconsistent concepts for which it is not obvious how to construct an inconsistent definition. Moreover, given that all the inconsistent concepts Yablo

[31] Yablo (1993a, 1993b).

considers are inconsistent by definition, all these concepts are essentially inconsistent. However, **up above** is an *empirically inconsistent concept*. It is inconsistent, in part, because of the environment in which it is used. If the Higherians all lived on a flat surface of a planet, then 'up above' would not be inconsistent. The moral is that a theory of inconsistent concepts must treat empirically inconsistent ones along with the logically inconsistent ones. Yablo's account works only for the latter. Therefore, it fails to satisfy the fourth condition on theories of inconsistent concepts.

E.8 DIALETHEISM

Dialetheism is the view that some truth bearers are both true and false. As such, it is not a theory of inconsistent concepts. However, it has been used in the construction of theories of inconsistent concepts. In fact, if one accepts dialetheism, then one can treat the set of meaning-constitutive principles for an inconsistent concept as a theory of that concept. Of course, the theory will be inconsistent, but according to the dialetheist, some inconsistent theories are acceptable. To accompany a dialetheist theory of inconsistent concepts, the dialetheist endorses a paraconsistent logic. The most common one is LP. On LP, sentences have one of three truth-values: true, false or both true and false (glut). True and glut are designated. LP is a *paraconsistent* logic, which means that it is not the case that everything follows from a contradiction in LP. However, LP is not trivial—it is not the case that all sentences are theorems. Thus, one can use it to draw a substantive distinction between theorems and non-theorems and between valid arguments and invalid arguments.

On a dialetheic theory of **up above**, some of the Higherians' sentences that contain 'up above' are gluts. Indeed, because the constitutive principles for 'up above' conflict in many cases, most of the sentences containing 'up above' that are considered true by the Higherians are gluts on the dialetheic theory. The dialetheist endorses LP as a logic for the Higherians' language; it provides a standard by which one can assess their arguments for validity.

Does the dialetheic theory of inconsistent concepts satisfy the conditions on theories of inconsistent concepts? It seems to me that the principle of mono-aletheism (i.e., no truth-bearer is both true and false) is constitutive of our concept of truth. Thus, it seems to me that one can reject dialetheism and the dialetheic theory of inconsistent concepts without giving any further justification.[32] However, it has another problem as well. It is a meaning-constitutive theory, which means that it cannot distinguish between concept possession and concept employment. Anyone who accepts the dialetheic theory employs an inconsistent concept and, thus, accepts an inconsistent set of principles. Thus, it fails to satisfy condition three.

E.9 FRAMES

In this section, I discuss Gupta's suggestion for interpreting the Higherians. His suggestion requires a distinction between *absolute* and *effective* features of the Higherians' discursive practice. In particular, their linguistic expressions, mental states, and performances have both absolute pragmatic and semantic features and effective pragmatic and semantic features. For example, 'up above' has an absolute meaning and an effective meaning, a sentence containing 'up above' has an absolute truth-value and an effective truth-value, and an utterance of such a

[32] See Lewis (1982: 101) for a similar view.

sentence has an absolute force and an effective force. The difference between an absolute feature and an effective feature is that a frame figures in the determination of an effective feature, but not for an absolute feature.

Gupta's idea is that a person who employs an inconsistent concept without knowing it is inconsistent privileges certain constitutive principles in certain situations. A *frame* is a way of privileging certain constitutive principles over others when employing an inconsistent concept. In the case of 'up above', the Higherians privilege the perceptual criterion in some cases and they privilege the conceptual criterion in others. Call the former the *perceptual frame* and the latter the *conceptual frame*. One uses a frame to determine the effective semantic and pragmatic features of the Higherians' linguistic expressions, mental states, and performances.

It is common to assume that the meaning of a sentence and the context in which it is uttered determine its content, and that the content of a sentence and a possible world determine its truth-value in that world. These are absolute features of the sentence. Likewise, the meaning of a sentence, the context in which it is uttered, and a frame determine its effective content, and the effective content of a sentence and a possible world determine its effective truth-value in that world. One can draw similar distinctions for other semantic and pragmatic concepts.

Gupta claims that a frame is determined by the practice of those who employ the concept in question. Some employments of the concept might not have a frame associated with them. In addition, he cautions against thinking of frames as contexts or as determined by the rules of a language. Rather, a frame can be in effect in many different contexts, and frames ways of interpreting the rules of a language. Beyond these remarks, Gupta says very little on how to interpret the Higherians.

Because Gupta's theory is currently inchoate, it is hard to say much about it. One thing should be clear: Gupta's theory is not a theory of inconsistent concepts—he offers no account of the absolute features of the Higherians' linguistic expressions, mental states, and performances. That is *not* a criticism of Gupta. Indeed, I claim that there is an important place in a theory of inconsistent concepts for a theory of the sort Gupta presents. Gupta's theory is a theory of how those who employ an inconsistent concept actually use it. It seems to me that employers of an inconsistent concept do privilege certain constitutive principles when employing it in a given situation. Gupta gives us the tools to makes sense of this behavior. However, his account is not a theory of inconsistent concepts. Thus, Gupta's theory should not be thought of as a competitor to the other theories I discuss in this paper.

One might be tempted to assume that the effective semantic and pragmatic features just are their absolute semantic and pragmatic features; then one would have a genuine theory of inconsistent concepts based on Gupta's theory. I agree—one would have a genuine theory of inconsistent concepts. However, it would be an inadequate theory. The first problem is that this theory implies that inconsistent concepts are context dependent. For example, one can find two objects, A and B, such that in one context the perceptual frame is effective and, consequently, <A, B> is in the extension of 'up above', but the conceptual frame is effective in a different context, which renders <A, B> a member of its anti-extension (assume that A and B are in the same positions in each context). Thus, the extension of 'up above' depends on the frame that is effective, and so, on the context in which the utterance occurs. We have already seen that context dependence theories of inconsistent concepts are inadequate because one can introduce concepts that are inconsistent and context-invariant.

A second problem with this theory (again, it is not Gupta's theory—it is one way of augmenting Gupta's theory) is that it is incomplete. In order to determine the effective content of a sentence, one must know its meaning, the context in which it is uttered, and the relevant frame. Which type of meaning? Is it the effective meaning or the absolute meaning? If it is the effective meaning, then how do we determine that? Gupta's theory gives us no account of it. If it is the absolute meaning, then Gupta's theory depends on a theory of inconsistent concepts, which would specify the absolute meaning of the sentence. Therefore, altering Gupta's theory in this way does not constitute an adequate theory of inconsistent concepts.

I want to return to the theory Gupta actually presents and discuss several issues surrounding it. First the notion of a frame is imprecise. It seems to me that we should think of a frame as an ordering of the constitutive principles for the concept in question. In particular, a *frame* is a function from the set of constitutive principles to an initial segment of the ordinals. A constitutive principle with a lower ordinal trumps one assigned a higher ordinal if there is a conflict. The function can assign two constitutive principles the same ordinal if and only if they never conflict. For example, the perceptual frame assigns the perceptual criterion 0 and the conceptual criterion 1; the conceptual frame assigns the conceptual criterion 0 and the perceptual criterion 1. If the perceptual frame is in effect, then the perceptual criterion determines the effective extension of 'up above', and the conceptual criterion is ignored if the two conflict.[33]

Another issue is that sentences and arguments that are evaluated with multiple frames might pose a problem. For example, assume that a Higherian, Hiram, asserts the following sentence: if A is up above B, then it is not the case that B is up above A. Call this sentence *p*.

[33] I suggest that we could abandon the idea that there are multiple discrete frames that are in play for a particular inconsistent concept and treat *the frame* as a function from the product of the set of constitutive principles and the set of contexts to an initial segment of the ordinals. On this suggestion, the frame determines the way the principles are weighted in every context.

Assume that ‘A’ and ‘B’ are functioning as names in p, that A is perceptually up above B, and that B is conceptually up above A (e.g., A and B are on the other side of the planet from Standard Up). Assume as well that the perceptual frame is in effect for the first occurrence of ‘up above’ in p and that the conceptual frame is in effect for the second occurrence of ‘up above’ in p. The antecedent of p is effectively true and the consequent is effectively false; hence, the conditional is effectively false. That is a strange consequence given that another constitutive principle of ‘up above’ should be that it is anti-symmetric. Thus, sentences evaluated from multiple frames can violate other constitutive principles of the concept in question.

A related example poses a problem for arguments whose sentences are evaluated from multiple frames:

(a) A is up above B.

∴ (b) A is up above B.

If A and B are located away from Standard Up, A is perceptually up above B, the perceptual criterion is in effect for (a), and the conceptual criterion is in effect for (b), then the argument is effectively invalid (i.e., (a) is effectively true and (b) is effectively false). Thus, if we use a logic to evaluate the Higherians’ arguments for validity that is based on the effective truth values of their sentences, then we treat them as if they are inferentially irrational (i.e., as if they cannot follow their own inference rules). Thus, it is similar to the context dependence theory and the ambiguity theory in this regard. We have to be careful if we use the effective features of the Higherians’ linguistic expressions in a logic or a semantic theory for ‘up above’.

I see at least three roles for Gupta’s theory in a theory of inconsistent concepts. First, it allows an interpreter of a discursive practice in which an inconsistent concept is employed to predict how the members of that community will use the concept. An interpreter can observe the members of the community long enough to determine the frame they use, then he can use it to explain their employments of the concept. He can use it to explain why they apply it and disapply it in the way they do, why they do not recognize that it is inconsistent, and why they do not accept contradictions by applying it. It explains all the semantic and pragmatic features those who employ the inconsistent concept attribute to the linguistic expressions, mental states, and performances associated with it. For example, it explains: (i) the truth-values they attribute to sentences that express the inconsistent concept (i.e., the effective truth-values), (ii) the validity-values they attribute to arguments containing such sentences (i.e., the effective validity-values), and (iii) the forces they attribute to utterances of such sentences (i.e., the effective force-values).[34]

Second, the theory can be assimilated into a pragmatic theory for inconsistent concepts. In Chapter Six, I present such a scorekeeping-based pragmatic theory for inconsistent concepts. One issue I do not discuss is how someone who knows that a concept is inconsistent should keep track of how the people who employ it keep score on one another. It seems to me that Gupta's theory would work well for this purpose because it explains the semantic and pragmatic features attributed by the employers of the inconsistent concept. For example, a person who knows that 'up above' is inconsistent and is keeping score on a group of Higherians can use Gupta's theory to keep track of the commitments and entitlements the Higherians attribute to each other when using 'up above'.

Third, Gupta's theory can be used as part of an account of how someone who employs an inconsistent concept comes to discover that it is inconsistent. For example, assume that Hiram sees two immobile objects, A and B, such that A is perceptually up above B, but A is not

[34] It is interesting that, because the Higherians evaluate one another's arguments based on effective truth-values, they treat one another as inferentially irrational without knowing it.

conceptually above B. Hiram asserts ‘A is up above B’ and he writes this down. It seems that the perceptual frame is in effect for this assertion. Thus, the sentence asserted is effectively true. Sometime later he has forgotten his report about A and B, he cannot find the paper on which he wrote it down, and he needs to determine whether A is up above B; however, at this time B is obscured. Instead of using the perceptual criterion, Hiram decides to use the conceptual criterion to determine whether A is up above B. He does so and discovers that the ray connecting A and B is not parallel to Standard Up. He asserts ‘A is not up above B’, and writes this report down. It seems that the conceptual frame is in effect for this assertion. Thus, the sentence asserted is effectively true. At some later time, he finds both written reports. He knows that A and B have not moved, so he is confused. At one time he wrote that A is up above B and at another he wrote that A is not up above B. He wonders what happened. He decides to use the perceptual criterion and the conceptual criterion at the same time. He is astonished to discover that they deliver different results. He uses the perceptual criterion and asserts ‘A is up above B’; he uses the conceptual criterion and asserts ‘A is not up above B’. Each of these is effectively true. Hiram then infers ‘A is up above B and A is not up above B’. Again, this sentence is effectively true. However, he has derived a contradiction. Of course, he knows that contradictions cannot be true. He is perplexed. He has been properly using the concept according to its constitutive principles, but he has derived a contradiction from intuitively correct assumptions via intuitively correct inference rules. He decides after checking that he has not made any errors that he has discovered a paradox. Perhaps he calls it the *higher paradox*. It seems to me that neither the perceptual frame nor the conceptual frame is in effect from this point on. Instead, we should think of this as a “conceptual time out.” He is no longer attributing truth-values to the sentences in which ‘up above’ occurs and he no longer accepts either of them. He has decided to try to figure out what has gone wrong before he continues using ‘up above’ in the unselfconscious way he did prior to his discovery. He is now engaged in an investigation into his own conceptual repertoire.

We can assume that after some time, he discovers that the conceptual criterion and the perceptual criterion deliver conflicting results for the vast majority of cases. This conclusion probably comes after years of suggestions about the “mistakes” he is making in the derivation of the higher paradox (e.g., the sentences involved are meaningless, ‘up above’ is partially defined, the proper logic does not validate his reasoning). Gupta’s theory helps us understand this process. In ordinary cases, a frame is in effect and the person employing the concept does so without a second thought about the concept itself. The theory posits effective semantic and pragmatic features to explain the person’s use of the concept. These effective semantic and pragmatic features might eventually lead users of the concept to discover a paradox associated with it; the paradox is often the derivation of an effectively true contradiction. Once this occurs, they begin looking into their employment of the concept in a new way. We can think of this occurring without a frame in effect at all. Eventually they discover that the problem lies not in how they have been using the concept, in the intuitions they have about it, or the inference rules they use for reasoning—the problem is their concept itself.[35]

There is, of course, much more to be said about this topic, but it is beyond the scope of this paper. Furthermore, there is much more that needs to be said about Gupta’s theory before it will fill these three roles in a theory of inconsistent concepts; it will have to wait for some other occasion. In summary, Gupta’s theory is not a theory of inconsistent concepts; rather, it is a theory of how inconsistent concepts are actually employed by those who are ignorant of their inconsistency. There is an important place for this theory in an acceptable theory of inconsistent concepts.

[35] See Camp (2002: chs. 14-16) for discussion of how one discovers that a concept one employs is inconsistent.

E.10 Confusion

In this final section, I discuss the theory of inconsistent concepts I endorse. On this theory, an inconsistent concept is one whose constitutive principles are inconsistent. The central claim of this theory is that inconsistent concepts are confused concepts. That is, for each inconsistent concept, there is a set of component concepts that play two roles. First, they are used in the logic, the pragmatic theory, and the semantic theory for inconsistent concepts; second, they serve as replacements for the inconsistent concept.

When considering an inconsistent concept, it is essential to distinguish between several sets of rules for using it. First, there are the inconsistent rules that are constitutive of the concept. Those who employ the concept try to follow these rules. Second, there are the rules stipulated by the logic, the pragmatic theory, and the semantic theory for inconsistent concepts. An interpreter who knows the concept is inconsistent treats those who employ it as if they are bound by these rules. Third, there are the rules stipulating that the concept should not be used at all. Those who know that it is inconsistent are bound by these rules.

It is essential to distinguish between an inconsistent concept's application set, its extension, its disapplication set, its anti-extension, its range of inapplicability, and its non-extension. A concept's *application set* includes all the items to which applies, its *disapplication set* contains all the items to which it disapplies, and its *range of inapplicability* consists of all the items to which it neither applies nor disapplies (all three are determined by its constitutive principles). (The union of its application set and disapplication set is its *range of applicability*.) For acceptable concepts, the extension and the application set are identical, the anti-extension and the disapplication set are identical, and the non-extension and the range of inapplicability are identical. An inconsistent concept has an empty extension and anti-extension, but its application set and a disapplication set need not be. An inconsistent concept can be application-inconsistent or range-inconsistent (or both). A concept is *application-inconsistent* if and only if its application set and disapplication set are not disjoint. A concept is *range-inconsistent* if and only if its range of applicability and range of inapplicability are not disjoint.

The theory of inconsistent concepts I offer has a logic, a pragmatic theory, and a semantic theory for inconsistent concepts. The logic appropriate for an inconsistent concept depends on its components. If it is completely defined and has n components, then an n-component logic is appropriate. If it is partially defined and has n components, then a partial n-component logic is appropriate. Both n-component logics and partial n-component logics are relevance logics. The pragmatic theory is a scorekeeping theory—it specifies how those who know the concept is inconsistent keep score on those who employ it. The semantic theory for inconsistent concepts is an inferential role theory—it specifies the inferential roles of the sentences that express the inconsistent concept (the inferential role of a sentence includes its role in perception and action as well). These three theories are used to interpret those who employ inconsistent concepts. In addition, I endorse the replacement policy for handling cases of conceptual inconsistency: inconsistent concepts should be replaced with consistent ones. Thus, the theory I offer satisfies the first condition (i.e., it is a genuine theory of inconsistent concepts).

The logic I advocate is inferentially charitable. In particular, the logic implies that the employer of the confused concept is capable of following inference rules she accepts, that she is capable of following the rules of this logic, and that she is motivated to follow the rules of this logic. Of course, the logic treats certain classically valid inference rules as invalid; thus, it implies that the person who employs an inconsistent concept has accepted the wrong inference rules. However, this consequence is different from the claim that the person who employs an inconsistent concept is incapable of following the inference rules he accepts. Thus, the theory of inconsistent concepts satisfies the second condition (i.e., it should be inferentially charitable).

I say very little about concept possession and concept employment. I can say that a person *employs* a particular concept if and only if he is disposed to apply it or disapply it in some circumstances (after due reflection). A person *applies* (*disapplies*) a concept X if and only if he is disposed to utter a sentence of the form ⟨α is X⟩ (⟨α is not X⟩) or he accepts the belief expressed by this sentence (after due reflection). I favor an account of concept possession that is based on understanding. A person *possesses* a concept X if and only if he can understand some sentences that express X. Therefore, although I endorse a theory on which inconsistent concepts have constitutive principles, it does not fall prey to the criticism of cognitive meaning-constitutive theories I presented in E.2.2. On the theory I endorse, the constitutive principles are involved in the *employment* of the concept, not in its *possession*. When a person employs a concept, she commits herself to obeying its constitutive principles. However, one can possess a concept without employing it.

On the theory I offer, one can possess an inconsistent concept, one can attribute it to someone else, and one can use the logic, the pragmatic theory, and the semantic theory for inconsistent concepts without employing the inconsistent concept in question. Indeed, on this account, one attributes an inconsistent concept to someone if one deems it appropriate to use the logic, the pragmatic theory, and the semantic theory to interpret that person. However, the logic, the pragmatic theory, and the semantic theory appeal to the replacement concepts, not to the inconsistent concept. Thus, the theory I offer satisfies condition three (i.e., it permits a distinction between concept possession and concept employment).

Finally, this theory of inconsistent concepts allows a wide range of views on the constitutive principles for an inconsistent concept. It permits the rules to appeal to empirical matters, which might have aspects unknown to the possessor of the concept. Thus, it applies to both intrinsically inconsistent concepts and empirically inconsistent concepts. Hence, it satisfies the fourth condition.

To illustrate the theory, consider how it applies in the case of the Higherians. **Up above** is an inconsistent concept. The following definition seems to capture it:

> (5a) 'up above' applies to <x,y> if x and y are observable, x and y are within the borders of the Higherians' nation, and either x and y satisfy a particular perceptually distinguishable relation, *perceptually-up-above*, or the ray that connects x and y is parallel to a particular ray, *Standard Up*, and y is closer to the surface of the Higherians' planet than x.
>
> (5b) 'up above' disapplies to <x,y> if x and y are observable, x and y are within the borders of the Higherians' nation, and either x and y do not satisfy the *perceptually-up-above* relation or either the ray that connects x and y is not parallel to a particular ray, *Standard Up*, or y is not closer to the surface of the Higherians' planet than x.
>
> (5c) 'up above' is inapplicable to <x,y> if either x or y is not an observable object, or x or y are not within the borders of the Higherians' nation.

Note that **up above** is both inconsistent and partial. It seems to me that when trying to decide on the components of such a concept, one should first decide whether the concept is application-inconsistent, range-inconsistent, or both. If the concept is merely application-inconsistent, then it seems best to choose components that all have the same range of applicability as the original. **Up above** seems to be merely application-inconsistent.

The most natural way to construe **up above** as a confused concept is to think of it as the fusion of the perceptual component and the conceptual component. The perceptual component is based on the relation, *perceptually-up-above*, which applies to pairs of objects that can be observed simultaneously. To avoid the problems associated with response-dependent concepts, I will just assume that it is possible to describe the way in which two objects, one of which is perceptually up above the other, stimulate an appropriately placed observer's retinas. The conceptual component, I call *conceptually-up-above* is the relation that is based on the connecting ray being parallel to Standard Up. The following definitions should clarify these two concepts:

(6) x is *perceptually up above* y if and only if x and y are both observable objects, x and y are within the borders of the Higherians' nation, x and y are observable simultaneously, and x and y would stimulate an appropriately placed normal observer in normal conditions in way P.

(7) x is *conceptually up above* y if and only if x and y are both observable objects, x and y are within the borders of the Higherians' nation, the ray connecting them is parallel to Standard Up, and y is closer to the surface of the Higherians' planet than x.

These two concepts are both consistent and will serve as the components in my explanation of **up above**.

An interpreter of the Higherians (call her *Doris*) must have access to someone who is able to determine, for any two objects, whether one is conceptually up above the other and whether one is perceptually up above the other. The interpretation begins with an account of validity. Given that 'up above' is partially defined and has two components, a partial 2-component logic is appropriate. For any argument that contains an occurrence of 'up above', Doris should use the experts to determine the possible assignments of semantic values to the sentences containing 'up above'. Dorise then uses the partial 2-component logic to determine whether the argument in question is valid. (See Chapter Six for details.)

Doris also uses the pragmatic theory for inconsistent concepts in her interpretation of the Higherians. Doris should first refrain from attributing a truth-value to sentences that contain 'up above'. She should then decide whether the person who uttered it is entitled to it. The scorekeeping pragmatics can serve that purpose here. If Doris determines that the person who uttered the sentence in question is not entitled to it by any other means, she should determine whether he is inferentially entitled to it. If it follows from a claim to which he is entitled by the partial 2-component logic, then he is entitled to it; if not then he is not entitled to it. As I have argued, none of the sentences containing 'up above' have truth values, and the notion of warrant involved here is an internal one. There is a sense of 'warrant' in which any assertion of a sentence that contains 'up above' is unwarranted because the sentence expresses an inconsistent concept. This notion of warrant is not he one the pragmatic theory tracks. Instead, the pragmatic theory implies that some assertions of sentences containing 'up above' are warranted in the sense that, *ceteris paribus*, acting on them is likely to lead to the satisfaction of the Higherians' desires.

Finally, Doris should determine the freestanding inferential content and the freestanding assertional content of the sentences in which 'up above' occurs by using the semantic theory for inconsistent concepts. This theory works together with the logic and the pragmatic theory to explain why the Higherians' sentences that contain 'up above' have the contents they have.

E.11 CONCLUSION

In section two, I presented an example of an inconsistent concept and four conditions on a theory of inconsistent concepts. I then discussed nine theories of inconsistent concepts: the context dependence theory, the indirect context dependence theory, the disambiguation theory, Field's theory, Eklund's theory, Yablo's theory, the dialetheic theory, Gupta's theory, and my theory. The context dependence theory, the indirect context dependence theory, and Gupta's theory fail to satisfy the first condition (i.e., they are not genuine theories of inconsistent concepts). However, there is an important place for Gupta's theory in a genuine theory of inconsistent concepts. The disambiguation theory fails to satisfy the second condition (i.e., it implies that the employers of inconsistent concepts are irrational). Field's theory, Eklund's theory, and the dialetheic theory fail to satisfy the third condition (i.e., they do not permit a distinction between concept possession and concept employment). Yablo's theory fails to satisfy the fourth condition (i.e., it does not apply to empirically inconsistent concepts). The confusion-based theory of inconsistent concepts satisfies all four conditions: it treats inconsistent concepts as genuinely inconsistent, it is inferentially charitable, it respects the difference between possession and employment, and it applies to both essentially inconsistent concepts and empirically inconsistent concepts.

BIBLIOGRAPHY

Akiba, Ken. (2002a). "A Deflationist Approach to Indeterminacy and Vagueness," *Philosophical Studies* 107: 69-86.

———. (2002b). "Can Deflationism Allow for Hidden Indeterminacy?" *Pacific Philosophical Quarterly* 83: 223-234.

Allen, Martin. (Forthcoming). "A Paraconsistent-Preservationist Treatment of a Common Confusion Concerning Predicate Extensions."

Anderson, Alan R. and Belnap, Nuel. (1975). *Entailment: The Logic of Relevance and Necessity*, vol. 1. Princeton: Princeton University Press.

Anscombe, G. E. M. (1957). *Intention*. Ithaca, NY: Cornell University Press.

Antonelli, G. Aldo. (1994). "The Complexity of Revision," *Notre Dame Journal of Formal Logic* 35: 67-72.

———. (1996). Review of Simmons (1993), *Notre Dame Journal of Formal Logic* 37: 152-159.

———. (2000). "Virtuous Circles: From Fixed Points to Revision Rules," in Chapuis and Gupta (2000).

Armour-Garb, Bradley. (2001). "Deflationism and the Meaningless Strategy," *Analysis* 61: 280-289.

Armour-Garb, Bradley and Beall, J. C. (2001). "Can Deflationists be Dialetheists?" *Journal of Philosophical Logic* 30: 593-608.

Azzouni, Jodi. (2001). "Truth via Anaphorically Unrestricted Quantifiers," *Journal of Philosophical Logic* 30: 329-354.

Barwise, Jon, and Etchemendy, John. (1987). *The Liar: An Essay on Truth and Circularity*. Oxford: Oxford University Press.

Beall, JC. (2000). "Minimalism, Gaps, and the Holton Conditional," *Analysis* 60: 340-351.

———. (2001). "A Neglected Deflationary Approach to the Liar," *Analysis* 61: 129-136.

———. (2002). "Deflationism and Gaps: Untying 'Not's in the Debate," *Analysis* 62: 299-305.

———. (ed.). (2003). *Liars and Heaps: New Essays on Paradox*. Oxford: Clarendon Press.

Belnap, Nuel. (1976). "How a computer should think," in *Contemporary Aspects of Philosophy*, G. Ryle (ed.), London: Oriel Press.

———. (1977). "A Useful Four-Valued Logic," in *Modern Uses of Multiple-Valued Logic*, J. M. Dunn and G. Epstein (eds). Dordrecht: D. Reidel.

———. (1982). "Gupta's Rule of Revision Theory of Truth," *Journal of Philosophical Logic* 11: 103-116.

Blackburn, Simon. (1998). "Wittgenstein, Wright, Rorty, and Minimalism," *Mind* 107: 157-181.

Blamey, Stephen. (2002). "Partial Logic," in *Handbook of Philosophical Logic*, vol. 5, D. Gabbay and F. Guenthner (eds.), Dordrecht: Kluwer Academic Publishers.

Boghossian, Paul. (1989). "Content and Self-Knowledge," *Philosophical Topics* 17: 5-26.

———. (1990). "The Status of Content," *Philosophical Review* 99: 157-184.

———. (1996). "Analyticity Reconsidered," *Noûs* 30: 360-391.

———. (1997). "Analyticity," in *A Companion to the Philosophy of Language*, B. Hale and C. Wright (eds.), Oxford: Blackwell.

———. (2000), "Knowledge of Logic," in *New Essays on the A Priori*, P. Boghossian and C. Peacocke (eds.), Oxford: Oxford University Press.

Brandom, Robert. (1994). *Making It Explicit*. Cambridge: Harvard University Press.

———. (2002). "Explanatory vs. Expressive Deflationism about Truth," in Schantz (2002).

Brendel, Elke. (2000). "Circularity and the Debate Between Deflationist and Substantive Theories of Truth," in Chapuis and Gupta 2000.

Bromand, Joachim. (2002). "Why Paraconsistent Logic Can Only Tell Half the Truth," *Mind* 111: 741-749.

Brueckner, Anthony. (1999). "Difficulties in Generating Scepticism about Knowledge of Content," *Analysis* 59: 212-217.

———. (2000). "Ambiguity and Knowledge of Content," *Analysis* 60: 257-260.

Burge, Tyler. (1977). "Belief *De Re*," *The Journal of Philosophy* 74: 338-362.

———. (1979a). "Semantical Paradox," *The Journal of Philosophy* 76: 169-198.

———. (1979b). "Individualism and the Mental," *Midwest Studies in Philosophy* 4: 73-121.

———. (1982a). "Postscript to 'Semantical Paradox'," in Martin (1984).

———. (1982b). "The Liar Paradox: Tangles and Chains," *Philosophy Studies* 41: 353-366.

———. (1988). "Individualism and Self-Knowledge," *The Journal of Philosophy* 85: 649-663.

Burgess, John. (1986). "The Truth is Never Simple," *The Journal of Symbolic Logic* 51: 663-681.

———. (2002). "Is There a Problem about the Deflationary Theory of Truth?" in Halbach and Horsten 2002.

Camp Jr., Joseph L. (2002). *Confusion: A Study in the Theory of Knowledge.* Cambridge: Harvard University Press.

Cantini, Andrea. (1995). "Levels of Truth," *Notre Dame Journal of Formal Logic* 36: 185-213.

Cargile, James. (1986). Critical Notice of Martin (1984), *Mind* 95: 116-126.

Carruthers, Peter, and Smith, Peter K. (eds.). (1996). *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Chomsky, Noam. (1995). "Language and Nature," *Mind* 104: 1-61.

Chihara, Charles S. (1973). "A Diagnosis of the Liar and Other Semantical Vicious-Circle Paradoxes," *The Work of Bertrand Russell*, C. Roberts (ed.), London: Allen and Unwin.

———. (1979). "The Semantic Paradoxes: A Diagnostic Investigation," *Philosophical Review* 88: 590-618.

———. (1984a). "The Semantic Paradoxes: Some Second Thoughts," *Philosophical Studies* 45: 223-229.

———. (1984b). "Priest, the Liar, and Gödel," *Journal of Philosophical Logic* 13: 117-124.

Chapuis, André. (1996). "Alternative Revision Theories of Truth," *Journal of Philosophical Logic* 25: 399-423.

———. (2000). "Rationality and Circularity," in Chapuis and Gupta 2000.

Chapuis, André, and Gupta, Anil (eds.). (2000). *Circularity, Definition and Truth.* New Delhi: Indian Council of Philosophical Research.

Church, Alonzo. (1946). Review of: Koyré, Alexandre (1946), *The Journal of Symbolic Logic* 11: 131.

———. (1976). "Comparison of Russell's Resolution of the Semantical Antimonies with that of Tarski," *The Journal of Symbolic Logic* 41: 747-760.

Clark, Michael. (1997). "Truth and Success: Searle's Attack on Minimalism," *Analysis* 57: 205-209.

———. (1999). "Recalcitrant Variants of the Liar Paradox," *Analysis* 59: 117-126.

Cohen, L. Jonathan. (1957). "Can the Logic of Indirect Discourse be Formalised?" *The Journal of Symbolic Logic* 22: 225-232.

———. (1961). "Why Do Cretans Have to Say So Much?" *Philosophical Studies* 12: 72-78.

Cook, Roy T. (2002). "Counterintuitive Consequences of the Revision Theory of Truth," *Analysis* 62: 16-22.

David, Marian. (1989). "Truth, Eliminativism, and Disquotationalism," *Noûs* 23: 599-614.

———. (1994). *Correspondence and Disquotation: An Essay on the Nature of Truth.* Oxford: Oxford University Press.

Davidson, Donald. (1967). "Truth and Meaning," in Davidson (1984).

———. (1968). "On Saying That," in Davidson (1984).

———. (1973). "Radical Interpretation," in Davidson (1984).

———. (1974). "On the Very Idea of a Conceptual Scheme," in Davidson (1984).

———. (1979). "The Inscrutability of Reference," in Davidson (1984).

———. (1982). "Communication and Convention," in Davidson (1984).

———. (1984). *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

———. (1986). "A Nice Derangement of Epitaphs." in *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, E. LePore (ed.), Oxford: Blackwell.

———. (1987). "Knowing One's Own Mind," in Davidson (2001).

———. (1988). "The Myth of the Subjective," in Davidson (2001).

———. (1990). "The Structure and Content of Truth," *The Journal of Philosophy* 87: 279-328.

———. (1992). "The Second Person," in Davidson (2001).

———. (2001). *Subjective, Intersubjective, Objective*. Oxford: Oxford University Press.

———. (Forthcoming). *Truth and Predication*. Cambridge: Harvard University Press.

De Vidi, David, and Solomon, Graham. (1999). "Tarski on 'Essentially Richer' Metalanguages," *Journal of Philosophical Logic* 28: 1-28.

Dennett, Daniel. (1982). "Beyond Belief," in *Thought and Object: Essays on Intentionality*, A. Woodfield (ed.), Oxford: Oxford University Press.

———. (1987). *The Intentional Stance*. Cambridge: MIT Press.

DeRose, Keith (2002). "Assertion, Knowledge, and Context," *Philosophical Review* 111, 167-203.

———. (2004). "Single Scoreboard Semantics," *Philosophical Studies*, 119: 1-21.

Devitt, Michael. (1991). "Minimalist Truth: A Critical Notice of Paul Horwich's *Truth*," *Mind and Language* 6: 273-283.

Divers, John, and Miller, Alexander. (1994). "Why Expressivists about Value Should Not Love Minimalism about Truth," *Analysis* 54: 12-19.

Dodd, Julian. (2002). "Truth," *Philosophical Books* 43: 279-291.

Dreier, James. (1996). "Expressivist Embedding and Minimal Truth," *Philosophical Studies* 83: 29-51.

Dummett, Michael. (1978). *Truth and Other Enigmas.* Cambridge: Harvard University Press.

———. (1991). *The Logical Basis of Metaphysics*. Cambridge: Harvard University Press.

Dunn, J. Michael. (1966). The Algebra of Intensional Logics. Ph.D. Dissertation, University of Pittsburgh.

Dunn, J. Michael and Restall, Greg. (2002). "Relevance Logic," in *Handbook of Philosophical Logic*, vol. 6, D. Gabbay and F. Guenthner (eds.), Dordrecht: Kluwer Academic Publishers.

Eklund, Matti. (2002). "Inconsistent Languages," *Philosophy and Phenomenological Research* 64: 251-275.

Engel, Pascal. (2002). *Truth.* Montreal: McGill-Queen's University Press.

Evans, Gareth. (1982). *The Varieties of Reference*. J. McDowell (ed.). Oxford: Oxford University Press.

Everett, Anthony. (1996). "A Dilemma For Priest's Dialetheism?" *Australasian Journal of Philosophy* 74: 657-668.

Feferman, Solomon. (1982). "Toward Useful Type-Free Theories, I," *The Journal of Symbolic Logic* 49: 75-111.

Feldman, Richard. (2004). "Comments on DeRose's 'Single Scoreboard Semantics'," *Philosophical Studies* 119: 23-33.

Field, Hartry. (1972). "Tarski's Theory of Truth," in Field (2001a).

———. (1973). "Theory Change and the Indeterminacy of Reference," in Field (2001a).

———. (1974). "Quine and the Correspondence Theory," in Field (2001a).

———. (1986). "The Deflationary Conception of Truth," *Fact, Science and Morality*, G. McDonald (ed.). Oxford: Blackwell Publishers.

———. (1994a). "Deflationist Views of Meaning and Content," in Field (2001a).

———. (1994b). "Disquotional Truth and Factually Defective Discourse," in Field (2001a).

———. (1998). "Some Thoughts on Radical Indeterminacy," in Field (2001a).

———. (2000). "Indeterminacy, Degree of Belief, and Excluded Middle," in Field (2001a).

———. (2001a). *Truth and the Absence of Fact*. Oxford: Oxford.

———. (2001b). "Attributions of Meaning and Content," in Field (2001a).

———. (2001c). Postscript to Field 1998, in Field (2001a).

———. (2001d). Postscript to Field 2000, in Field (2001a).

———. (2002). "Saving Truth Schema From Paradox," *Journal of Philosophical Logic* 31: 1-27.

———. (2003a). "A Revenge-Immune Solution to the Semantic Paradoxes," *Journal of Philosophical Logic* 32: 139-177.

———. (2003b) "The Semantic Paradoxes and the Paradoxes of Vagueness," in Beall (2003).

———. (2003c). "No Fact of the Matter," *Australasian Journal of Philosophy* 81: 457-480.

———. (2004). "The Consistency of the Naïve theory of Properties," *The Philosophical Quarterly* 54: 78-104

———. (Forthcoming a). "Is the Liar Sentence Both True and False?" in *Deflationism and Paradox*, JC Beall and B. Armour-Garb (eds.), Oxford: Oxford University Press.

———. (Forthcoming b). "Variations on a Theme by Yablo," in *Deflationism and Paradox*, JC Beall and B. Armour-Garb (eds.), Oxford: Oxford University Press.

———. (Forthcoming c). "Mathematical Undecidables, Metaphysical Realism, and Equivalent Descriptions," in *The Philosophy of Hilary Putnam, Library of Living Philosophers*. Forthcoming.

Fine, Kit. (1975). "Vagueness, Truth, and Logic," *Synthese* 54: 235-59.

———. (1988). "Semantics for Quantified Relevance Logic," Journal of Philosophical Logic 17: 27-59.

Fitch, Frederic B. (1946). "Self-Reference in Philosophy," *Mind* 55: 64-73.

———. (1964). "Universal Metalanguages for Philosophy," *Review of Metaphysics* 17: 396-402.

Fitting, Melvin. (1986). "Notes on the Mathematical Aspects of Kripke's Theory of Truth," *Notre Dame Journal of Formal Logic* 27: 75-88.

Fodor, Jerry A. (1964). "On Knowing What We Would Say," *Philosophical Review* 73: 198-212.

Fodor, Jerry A., and Lepore, Ernest. (1996). "What cannot be Evaluated cannot be Evaluated and it cannot be Supervalued Either," *The Journal of Philosophy* 93: 516-535.

Forbes, Graeme. (1996). "Substitutivity and the Coherence of Quantifying In," *Philosophical Review* 105: 337-372.

Gaifman, Haim. (1988). "Operational Pointer Semantics: Solution to Self-Referential Puzzles I," in *Theoretical Aspects of Reasoning about Knowledge*, M. Vardi (ed.), Los Angeles: Morgan Kaufmann.

———. (1992). "Pointers to Truth," *The Journal of Philosophy* 89: 223-261.

———. (2000). "Pointers to Proposition," in Chapuis and Gupta (2000).

Gauker, Christopher. (2001). "T-Schema Deflationism versus Gödel's First Incompleteness Theorem," *Analysis* 61: 129-136.

Gentzen, Gerhard. (1969). *Collected Papers of Gerhard Gentzen*. E. Szabo (ed.). Amsterdam: North-Holland.

Gibbard, Allan. (1990). *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press.

Gibbons, John. (1996). "Externalism and Knowledge of Content," *Philosophical Review* 105: 287-310.

Glanzberg, Michael. (2001). "The Liar in Context," *Philosophical Studies* 103: 217-251.

———. (2003). "Against Truth-Value Gaps," in Beall (2003).

———. (2004). "A Contextual-Hierarchical Approach to Truth and the Liar Paradox," *Journal of Philosophical Logic* 33: 27-88.

———. (2005). "Truth, Reflection, and Hierarchies," *Synthese* 142: 289-315.

Goddard, Leonard. (1984). "The Nature of Reflexive Paradoxes: Part II," *Notre Dame Journal of Formal Logic* 25: 27-58.

Goddard, Leonard, and Johnston, Mark. (1983). "The Nature of Reflexive Paradoxes: Part I," *Notre Dame Journal of Formal Logic* 24: 491-508.

Gödel, Kurt. (1931), " On Formally Undecidable Propositions of Principia Mathematica and Related Systems I," in *From Frege to Gödel*, Van Heijenoort (ed.), Cambridge: Harvard University Press.

Goldberg, Sanford. (1997). "Self-Ascription, Self-Knowledge, and the Memory Argument," *Analysis* 57: 211-219.

———. (1999). "Word-Ambiguity, World-Switching, and Knowledge of Content: Reply to Brueckner," *Analysis* 59: 212-217.

———. (2000). "Word-Ambiguity, World-switching, and Semantic Intentions," *Analysis* 60: 260-264.

Goldstein, Laurence. (1992). "'This Statement is not True' is not True," *Analysis* 52: 1-5.

———. (1999). "A Unified Solution to Some Paradoxes," *Proceedings of the Aristotelian Society* 100: 53-74.

———. (2001). “Truth-Bearers and the Liar: A Reply to Alan Weir,” *Analysis* 61: 115-126.

Goldstein, Laurence, and Goddard, Leonard. (1980). “Strengthened Paradoxes,” *Australasian Journal of Philosophy* 58: 211-221.

Goodman, Nelson. (1955). *Fact, Fiction, and Forecast*. Cambridge: Harvard University Press.

Grattan-Guinness, I. (1998). “Structural Similarity of Structuralism? Comments on Priest’s Analysis of the Paradoxes of Self-Reference,” *Mind* 107: 823-834.

Grim, Patrick. (1991). *The Incomplete Universe: Totality, Knowledge, and Truth.* Cambridge: The MIT Press.

Grover, Dorothy. (1976). “‘This is False’ on the Presentential Theory,” *Analysis* 36: 80-83.

———. (1977). “Inheritors and Paradox,” *The Journal of Philosophy* 74: 500-604.

———. (1992). *A Prosentential Theory of Truth*. Princeton: Princeton University Press.

Grover, Dorothy, Camp, Joseph, and Belnap, Nuel. (1975). “A Prosentential Theory of Truth,” in Grover (1992).

Gupta, Anil. (1982). “Truth and Paradox,” *Journal of Philosophical Logic* 11: 1-60.

———. (1983). “Postscript to ‘Truth and Paradox’,” in Martin (1984).

———. (1989). “The Liar: An Essay on Truth and Circularity,” *Philosophy of Science* 56: 697-709.

———. (1990). “Two Theorems Concerning Stability,” in *Truth or Consequences*, J. M. Dunn and A. Gupta (eds.), Dordrecht: Kluwer Academic Publishers.

———. (1993a). “A Critique of Deflationism,” *Philosophical Topics* 21: 57-81.

———. (1993b). “Minimalism,” *Philosophical Perspecitives* 7: 359-369.

———. (1997). “Definition and Revision: A Response to McGee and Martin,” *Philosophical Issues* 8: 419-443.

———. (1999). “Meaning and Misconceptions,” in *Language, Logic, and Concepts*. R. Jackendoff, P. Bloom and K. Wynn (eds.), Cambridge: MIT Press.

———. (2000). “On Circular Concepts,” in Chapuis and Gupta (2000).

———. (2001). “Truth,” in *The Blackwell Guide to Philosophical Logic,* L. Goble (ed.), Oxford: Blackwell.

———. (2002). “Partially Defined Predicates and Semantic Pathology,” *Philosophy and Phenomenological Research* 65: 402-409.

———. (Forthcoming). “Do the Paradoxes Pose a Special Problem for Deflationism?” in *Deflationism and Paradox*, JC Beall and B. Armour-Garb (eds.), Oxford: Oxford University Press.

Gupta, Anil, and Belnap, Nuel. (1993). *The Revision Theory of Truth.* Cambridge: MIT Press.

Gupta, Anil, and Martin, Robert L. (1984). “A Fixed Point Theorem for the Weak Kleene Valuation Scheme,” *Journal of Philosophical Logi*c 13: 131-135.

Halbach, Volker. (1995). “Tarski Hierarchies,” *Erkenntnis* 43: 339-367.

———. (1997). “Tarskian and Kripkean Truth,” *Journal of Philosophical Logic* 26: 69-80.

———. (1999). “Disquotationalism and Infinite Conjunctions,” *Mind* 108: 1-22.

———. (2000). “Disquotationalism Fortified,” in Chapuis and Gupta 2000.

———. (2002). “Modalized Disquotationalism,” in Halbach and Horsten 2002.

Halbach, Volker, and Horsten, Leon (eds.). (2002). *Principles of Truth*. Munich: Frankfurt am Main.

Hale, Bob and Wright, Crispin. (2000). “Implicit Definition and the A Priori,” in *New Essays on the A Priori*, P. Boghossian and C. Peacocke (eds.), Oxford: Oxford University Press.

Hardy, James. (1997). "Three Problems for the Singularity Theory of Truth," *Journal of Philosophical Logic* 26: 501-520.

Harman, Gilbert. (1986). *Change in View: Principles of Reasoning*. Cambridge: MIT Press.

———. (1995). "Rationality," in *Thinking: Invitation to Cognitive Science*, vol. 3, E. Smith and D. Osherson (eds.), Cambridge: MIT Press.

Hart, W. D. (1989). "Discussions for Anil Gupta," *Proceedings of the Aristotelian Society* 89: 161-165.

Hawthorne, John. (1990). "A Note on 'Languages and Language'," *Australasian Journal of Philosophy* 68:116- 119.

———. (2004). *Knowledge and Lotteries*. Oxford: Oxford University Press.

Hazen, Allen. (1981). "Davis's Formulation of Kripke's Theory of Truth: A Correction," *Journal of Philosophical Logic* 10: 309-311.

———. (1987). "Contra Buridanum," *Canadian Journal of Philosophy* 17: 875-880.

Heal, Jane. (2001). "On First-Person Authority," *Proceedings of the Aristotelian Society* : 1-19

Herzberger, Hans G. (1966). "The Logical Consistency of Language," in *Language and Learning*. J. A. Emig, J. T. Fleming, and H. M. Popp (eds.), New York: Harcourt, Brace and World: 250-263.

———. (1970a). "Paradoxes of Grounding in Semantics," *The Journal of Philosophy* 67: 145-167.

———. (1970b). "Truth and Modality in Semantically Closed Languages," in Martin (1970).

———. (1981). "New Paradoxes for Old," *Proceedings of the Aristotelian Society 81*: 109-123.

———. (1982a). "Naïve Semantics and the Liar Paradox," *The Journal of Philosophy* 79: 479-497.

———. (1982b). "Notes on Naïve Semantics," *Journal of Philosophical Logic* 11: 61-102.

Hinckfuss, Ian. (1991). "Pro Buridano: Contra Hazenum," *Canadian Journal of Philosophy* 21: 389-398.

Hintikka, Jaakko. (1962). *Knowledge and Belief*. Ithaca, NY: Cornell University Press.

Hodges, Wilfred. (1986). "Truth in a Structure," *Proceedings of the Aristotelian Society* 86: 135-151.

Hohwy, Jokob. (2002). "Privileged Self-Knowledge and Externalism: A Contextualist Approach," *Pacific Philosophical Quarterly* 83: 235-252.

Holton, Richard. (2000). "Minimalism and Truth-Value Gaps," *Philosophical Studies* 97: 137-168.

Horn, Lawrence. (1989). *A Natural History of Negation*. Stanford: CSLI Publications.

Horwich, Paul. (1982). "Three Forms of Realism," *Synthese* 51: 181-201.

———. (1994). "The Essence of Expressivism," *Analysis* 54: 19-20.

———. (1998a). *Truth.* 2nd ed. Oxford: Clarendon Press.

———. (1998b). *Meaning*. Oxford: Oxford University Press.

———. (2001). "A Defense of Minimalism," *Synthese* 126: 149-165.

Huggett, W. J. (1958). "Paradox Lost," *Analysis* 19: 21-23.

Humberstone, Lloyd. (2000). "The Revival of Rejective Negation," *Journal of Philosophical Logic* 29: 331-381.

Jackson, Frank, Oppy, Graham, and Smith, Michael. (1994). "Minimalism and Truth Aptness," *Mind* 103: 287-302.

Jorgensen, Jorgen. (1953). "Some Reflections on Reflexivity," *Mind* 62: 289-300.

———. (1955). "On Kattsoff's Reflections on Jorgensen's Reflections in Reflexivity," *Mind* 64: 542.

Juhl, Cory F. (1997). "A Context-Sensitive Liar," *Analysis* 57: 202-204.

Kalderon, Mark. (1997). "The Transparency of Truth," *Mind* 106: 475-497.

Kaplan, David. (1969). "Quantifying In," in *Words and Objections*, D. Davidson and J. Hintikka (eds.), Reidel: Dordrecht.

———. (1973). "Bob and Carol and Ted and Alice," in *Approaches to Natural Language*, K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes (eds.), Dordrecht: Reidel.

———. (1986). "Opacity," in *The Philosophy of W. V. Quine*, *The Library of Living Philosophers vol. 18*, L. E. Hahn and P. A. Schilpp (eds.), LaSalle, IL: Open Court.

———. (1989). "Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals," in *Themes From Kaplan*, J. Almog, J. Perry, and H. K. Wettstein (eds.), Oxford: Oxford University Press.

———. (1990). "Words," *Proceedings of the Aristotelian Society Supplement* 64: 93–119.

Kattsoff, L. O. (1955). "Some Reflections on Jorgensen's Reflections in Reflexivity," *Mind* 64: 96-98.

Kearns, John. (1970). "Some Remarks Prompted by van Fraassen's Paper," in Martin (1970).

Keefe, Rosanna. (2000). "Supervaluationism and Validity," *Philosophical Topics* 28: 93-105.

Ketland, Jeffrey. (2000). "A Proof of the (Strengthened) Liar Formula in a Semantical Extension of Peano Arithmetic," *Analysis* 60: 1-4.

———. (2003). "Can a Many-Valued Language Functionally Represent its Own Semantics?" *Analysis* 63: 292-297.

Kijania-Placek, Katarzyna. (2002). "What Difference Does it Make: Three Truth-Values or Two Plus Gaps," *Erkenntnis* 56: 83-98.

King, Jeffrey. (1994). "Can Propositions be Naturalistically Acceptable?" *Midwest Studies in Philosophy* 19: 53-75.

———. (1995). "Structured Propositions and Complex Predicates," *Noûs* 29: 516-535.

Koyré, Alexandre. (1946). "The Liar," *Philosophical and Phenomenological Research* 6: 344-362.

Kobes, Bernard. (1996). "Mental Content and Hot Self-Knowledge," *Philosophical Topices* 24: 71-99.

Koons, Robert C. (1992). *Paradoxes of Belief and Strategic Rationality.* Cambridge: Cambridge University Press.

———. (1994). Review of Gupta and Belnap (1993), *Notre Dame Journal of Formal Logic* 35: 606-631.

———. (2000). "Circularity and Hierarchy," in Chapuis and Gupta (2000).

Korsgaard, Christine. (1996). *The Sources of Normativity*, Cambridge: Cambridge University Press.

Kraut, Robert. (1993). "Robust Deflationism," *Philosophical Review* 102: 247-263.

Kremer, Michael. (1988). "Kripke and the Logic of Truth," *Journal of Philosophical Logic* 17: 225-278.

———. (2002). "Intuitive Consequences of the Revision Theory of Truth," *Analysis* 62: 330-336.

Kremer, Michael and Kremer, Philip. (2003). "Some Supervaluation-Based Consequence Relations," *Journal of Philosophical Logic* 32: 225-244.

Kremer, Philip. (1993). "The Gupta-Belnap System $S^{\#}$ and S^{*} are Not Axiomatisable," *Notre Dame Journal of Formal Logic* 34: 583-596.

———. (2000). "On the 'Semantics' for Languages with Their Own Truth Predicates," in Chapuis and Gupta 2000.

Kripke, Saul. (1972). *Naming and Necessity*. Cambridge: Harvard University Press. 1980.

———. (1975). "Outline of a Theory of Truth," *The Journal of Philosophy* 72: 690-716.

———. (1977). "Speaker's Reference and Semantic Reference," in *Contemporary Perspectives in the Philosophy of Language*, P. French, T. Uehling, and H. Wettstein (eds.), Minneapolis: University of Minnesota Press, 1979.

———. (1979). "A Puzzle about Belief," in *Meaning and Use*, A. Margalit (ed.), Reidel: Dordrecht.

———. (1982). *Wittgenstein on Rules and Private Language*. Cambridge: Harvard University Press.

Künne, Wolfgang. (2002). "Disquotationalist Conceptions of Truth," in Schantz (2002).

———. (2003). *Conceptions of Truth*. Oxford: Clarendon Press.

Lance, Mark. (1996). "Quantification, Substitution, and Conceptual Content," *Noûs* 30: 481-507.

———. (1998). "Some Reflections on the Sport of Language," *Philosophical Perspectives* 12: 219-240.

———. (2001). "The Logical Structure of Linguistic Commitment III: Brandomian Scorekeeping and Incompatibility," *Journal of Philosophical Logic* 30: 439-464.

Laporte, Joseph. (2003). *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

Leeds, Stephen. (1978). "Theories of Reference and Truth," *Erkenntnis* 13: 111-129.

———. (1995). "Truth, Correspondence, and Success," *Philosophical Studies* 79: 1-36.

———. (1997). "Incommensurability and Vagueness," *Noûs* 31: 385-407.

———. (2000). "A Disquotationalist Looks at Vagueness," *Philosophical Topics* 28: 107-128.

Leitgeb, Hannes. (2001). "Truth as Translation – Part A," *Journal of Philosophical Logic* 30: 281-307.

Lewis, David. (1969). *Convention*. Cambridge: Harvard University Press.

———. (1975). "Languages and Language," in *Philosophical Papers* vol. 1, Oxford: Oxford University Press, 1983.

———. (1979a). "Scorekeeping in a Language Game," in *Philosophical Papers*, vol. 1. Oxford: Oxford University Press. 1983.

———. (1979b). "Counterfactual Dependence and Time's Arrow" in *Philosophical Papers*, vol. 2. Oxford: Oxford University Press. 1986

———. (1982). "Logic for Equivocators," in *Papers in Philosophical Logic*, Cambridge: Cambridge University Press, 1998.

———. (1994). "Humean Supervenience Debugged," in *Papers in Metaphysics and Epistemology*, Cambridge: Cambridge University Press, 1999.

Lynch, Michael P. (ed.). (2001). *The Nature of Truth.* Cambridge: The MIT Press.

Mackie, J. L. (1973). *Truth Probability and Paradox*: *Studies in Philosophical Logic.* Oxford: Clarendon Press.

Mares, Edwin. (2004a). *Relevant Logic: A Philosophical Interpretation*. Cambridge: Cambridge University Press.

———. (2004b). "'Four-Valued' Semantics for the Relevant Logic R," *Journal of Philosophical Logic* 33: 327-341.

Martin, Donald A. (1997). "Revision and its Rivals," *Philosophical Issues* 8: 407-418.

Martin, Robert L. (1967). "Toward a Solution to the Liar Paradox," *Philosophical Review* 76: 279-311.

———. (ed.). (1970). *The Paradox of the Liar.* New Haven: Yale University Press.

———. (1976). "Are Natural Languages Universal?" *Synthese* 32: 271-291.

———. (1977). "On Puzzling Classical Validity," *Philosophical Review* 86: 454-473.

———. (ed.). (1984). *Recent Essays on Truth and the Liar Paradox*. Oxford: Clarendon Press.

Martinich, A. P. (1983). "A Pragmatic Solution to the Liar Paradox," *Philosophical Studies* 43: 63-67.

Mates, Benson. (1981). "Two Antinomies," in *Skeptical Essays*, Chicago: Chicago University Press.

Maudlin, Tim. (2004). *Truth and Paradox: Solving the Riddles*. Oxford: Oxford University Press.

McCarthy, Timothy. (1988). "Ungroundedness in Classical Languages," *Journal of Philosophical Logic* 17: 61-74.

———. (1985). "Abstraction and Definability in Semantically Closed Structures," *Journal of Philosophical Logic* 14: 255-266.

McDonald, Brian Edison. (2000). "On Meaningfulness and Truth," *Journal of Philosophical Logic* 29: 433-482.

McDowell, John. (1977). "On the Sense and Reference of a Proper Name," in *Meaning, Knowledge, and Reality*, Cambridge: Harvard University Press, 1998.

———. (1984). "De Re Senses," in *Meaning, Knowledge, and Reality*, Cambridge: Harvard University Press, 1998.

McGee, Vann. (1985). "How Truthlike Can a Predicate Be? A Negative Result," *Journal of Philosophical Logic* 14: 399-410.

———. (1989). "Applying Kripke's Theory of Truth," *The Journal of Philosophy* 86: 530-539.

———. (1991). *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth.* Cambridge: Hackett Publishing Company.

———. (1992). "Maximal Consistent Sets of Instances of Tarski's Schema (T)," *Journal of Philosophical Logic* 21: 235-241.

———. (1993). "A Semantic Conception of Truth?" *Philosophical Topics* 21: 83-111.

———. (1994). "Afterword: Truth and Paradox," *Basic Topics in the Philosophy of Language*, R. M. Harnish (ed.), Englewood Cliffs: Prentice Hall.

———. 1997). "Revision," *Philosophical Issues* 8: 387-406.

———. (2000). "The Analysis of 'x is true' as 'for any p, if x = 'p', then p," in Chapuis and Gupta (2000).

McGinn, Colin. (2000). *The Mysterious Flame: Conscious Minds in a Material World*. New York: Basic Books.

Michael, Emily. (1975). "Peirce's Paradoxical Solution to the Liar's Paradox," *Notre Dame Journal of Formal Logic* 16: 369-374.

Michael, Fred Seymour. (2002). "Entailment and Bivalence," *Journal of Philosophical Logic* 31: 289-300.

Mills, Andrew P. (1995). "Unsettled Problems With Vague Truth," *Canadian Journal of Philosophy* 25: 103-117.

Mills, Eugene. (1998). "A Simple Solution to the Liar," *Philosophical Studies* 89: 197-212.

Morreau, Michael. (1999). "Supervaluation can Leave Truth-Value Gaps After All," *The Journal of Philosophy* 99: 148-156.

Myhill, John. (1975). "Levels of Implication," in *The Logical Enterprise*, A. R. Anderson, R. B. Marcus, and R. M. Martin (eds.), London: Yale University Press.

Orilia, Francesco. (2000). "Belief Revision and the Aletheic Paradoxes," in Chapuis and Gupta (2000).

Parsons, Charles. (1974). "The Liar Paradox," *Journal of Philosophical Logic* 3: 381-412.

———. (1983). "Postscript to 'The Liar Paradox'," in *Mathematics in Philosophy: Selected Essays*, Ithaca: Cornell University Press.

Parsons, Terence. (1984). "Assertion, Denial, and the Liar Paradox," *Journal of Philosophical Logic* 13: 137-152.

———. (1990). "True Contradictions," *Canadian Journal of Philosophy* 20: 335-354.

Peacocke, Christopher. (1981). "The Theory of Meaning in Analytical Philosophy," in *Contemporary Philosophy*, vol. 1, G. Flöistad (ed.), The Hague: Nijhoff.

———. (1992). *A Study of Concepts*. Cambridge: MIT Press.

Pollock, John L. (1977). "The Liar Strikes Back," *The Journal of Philosophy* 74: 604-606.

Popper, Karl R. (1954). "Self-Reference and Meaning in Ordinary Language," *Mind* 63: 162-169.

Price, Huw. (1998). "Three Norms of Assertibility, or how the MOA Became Extinct," *Philosophical Perspectives* 12: 41-54.

Priest, Graham. (1979). "The Logic of Paradox," *Journal of Philosophical Logic* 8: 219-241.

———. (1984). "Logic of Paradox Revisited," *Journal of Philosophical Logic* 13: 153-179.

———. (1987). *In Contradiction: A Study of the Transconsistent.* Dordrecht: Martinus Nijhoff Publishers.

———. (1990). "Boolean Negation and All That," *Journal of Philosophical Logic* 19: 201-215.

———. (1991). "Minimally Inconsistent LP," *Studia-Logica* 50: 321-331.

———. (1992). Review of McGee (1991), *Mind* 101: 586-590.

———. (1993). "Another Disguise of the Same Fundamental Problems: Barwise and Etchemendy on the Liar," *Australasian Journal of Philosophy* 71: 60-69.

———. (1994). "The Structure of the Paradoxes of Self-Reference," *Mind* 103: 25-34.

———. (1995). "Gaps and Gluts: Reply to Parsons," *Canadian Journal of Philosophy* 25: 57-66.

———. (1998a). "What is so Bad About Contradictions," *The Journal of Philosophy* 95: 410-426.

———. (1998b). "The Import of Inclosure: Some Comments on Grattan-Guiness," *Mind* 107: 835-840.

———. (2000). "Truth and Contradiction," *The Philosophical Quarterly* 50: 305-319.

Prior, A. N. (1958). "Epimenides the Cretan," *The Journal of Symbolic Logic* 23: 261-266.

———. (1961). "On a Family of Paradoxes," *Notre Dame Journal of Formal Logic* 2: 16-32.

———. (1960). "The Runabout Inference Ticket," *Analysis* 21: 38-39.

Putnam, Hilary. (1975). "On the Meaning of 'Meaning'," in *Mind, Language and Reality*, Cambridge: Cambridge University Press,

———. (1978). *Meaning and the Moral Sciences*. Boston: Routledge and Kegan Paul.

———. (1985). “A Comparison of Something with Something Else,” in *Words and Life*, J. Conant (ed.), Cambridge: Harvard University Press.

Quine, W. V. (1956). “Quantifiers and Propositional Attitudes,” *The Journal of Philosophy* 53: 177-187.

———. (1960). *Word and Object*. Cambridge: MIT Press.

———. (1961). “The Ways of Paradox,” in *The Ways of Paradox and Other Essays*, Cambridge: Harvard University Press, 1966.

———. (1992). *The Pursuit of Truth*. Cambridge: Harvard University Press.

Rathjen, Michael. (1999). “The Realm of Ordinal Analysis,” in *Sets and Proofs*, B. Cooper and J. Truss (eds.), Cambridge: Cambridge University Press.

Reinhardt, William N. (1986). “Some Remarks on Extending and Interpreting Theories with a Partial Predicate for Truth,” *Journal of Philosophical Logic* 15: 219-251.

Resnik, Michael. (1990). “Immanent Truth,” *Mind* 99: 405-424.

Restall, Greg. (1995). “Four-Valued Semantics for Relevant Logics (and Some of Their Rivals),” *Journal of Philosophical Logic* 24: 139-160.

Richard, M. (1996). “Propositional Quantification,” in *Logic and Reality: Essays on the Legacy of Arthur Prior*, B. J. Copeland (ed.), Oxford: Oxford University Press.

Richards, Thomas J. (1967). “Self-Referential Paradoxes,” *Mind* 76: 387-403.

Rogers, Hartley. (1967). *Theory of Recursive Functions and Effective Computability*. McGraw-Hill.

Ross, Alf. (1969). “On Self-Reference and a Puzzle in Constitutional Law,” *Mind* 78: 1-24.

Rozeboom, William W. (1957). “Is Epimenides Still Lying?” *Analysis* 18: 105-113.

Ryle, Gilbert. (1951). “Heterologicality,” *Analysis* 11: 61-69.

Sawyer, Sarah. (1999). “My Language Disquotes,” *Analysis* 59: 206-211.

Santambrogio, Marco. (2002). “Belief and Translation,” *The Journal of Philosophy* 99: 624-647.

Schantz, Richard (ed.). (2002). *What is Truth?* Berlin: Walter de Gruyter.

Scharp, Kevin. (Forthcoming). “Scorekeeping in a Defective Language Game,” *Pragmatics and Cognition*.

Schiffer, Stephen. (1993). “Actual-Language Relations,” *Philosophical Perspectives* 7: 231-258.

———. (1998). “Two Issues of Vagueness,” *The Monist* 81: 193-214.

Segal, Gabriel. (2000). *A Slim Book about Narrow Content*. Cambridge: MIT Press.

Sellars, Wilfrid. (1954). “Some Reflections on Language Games,” in *Science, Perception and Reality*, Atascadero, CA: Ridgeview Press, 1964.

Setiya, Kieran. (2004). “Against Internalism,” *Noûs* 38: 266-298.

Setzer, Anton. (1999). “Ordinal Systems,” in *Sets and Proofs*, B. Cooper and J. Truss (eds.), Cambridge: Cambridge University Press.

Shapiro, Stewart. (2002). “Incompleteness and Inconsistency,” *Mind* 111: 817-832.

———. (2003). “The Guru, the Logician, and the Deflationist: Truth and Logical Consequence,” *Noûs* 37: 113-132.

Sheard, Michael. (1994). “A Guide to Truth Predicates in the Modern Era,” *The Journal of Symbolic Logic* 59: 1032-1054.

Simmons, Keith. (1990). “The Diagonal Argument and the Liar,” *Journal of Philosophical Logic* 19: 277-303.

———. (1993). *Universality and the Liar: An Essay on Truth and the Diagonal Argument.* Cambridge: Cambridge University Press.

———. (1994). “Paradoxes and Denotation,” *Philosophical Studies* 76: 71-106.

———. (1999). “Deflationary Truth and the Liar,” *Journal of Philosophical Logic* 28: 455-488.

———. (2000). “Three Paradoxes: Circles and Singularities,” in Chapuis and Gupta (2000).

———. (2003). “Reference and Paradox,” in Beall (2003).

Skinner, R. C. (1959). “The Paradox of the Liar,” *Mind* 68: 322-335.

Skyrms, Brian. (1970a). “Return of the Liar: Three-Valued Logic and the Concept of Truth,” *American Philosophical Quarterly* 7: 153- 161.

———. (1970b). “Notes on Quantification and Self-Reference,” in Martin (1970).

———. (1982). “Intensional Aspects of Semantical Self-Reference,” in Martin (1984).

———. (2000). “Truth Dynamics,” in Chapuis and Gupta (2000).

Smith, Michael. (1994). “Why Expressivists about Value should Love Minimalism about Truth,” *Analysis* 54: 1-12.

Smith, Nicholas J. J. (2000). “The Principle of Uniform Solution (of the Paradoxes of Self-Reference),” *Mind* 109: 117-122.

Soames, Scott. (1984). “What is a Theory of Truth?” *The Journal of Philosophy* 81: 411-429.

———. (1997). “The Truth about Deflationism,” in *Philosphical Issues*, vol 8, E. Villanueva (ed.), Atascadero, CA: Ridgeview Press.

———. (1999). *Understanding Truth.* Oxford: Oxford University Press.

———. (2002a). “Précis of *Understanding Truth*,” *Philosophy and Phenomenological Research* 65: 397-401.

———. (2002b). “Replies,” *Philosophy and Phenomenological Research* 65: 429-451.

———. (2002c). *Beyond Rigidity*. Oxford: Oxford University Press.

Stalnaker, Robert. (1970). “Pragmatics,” in Stalnaker (1999).

———. (1974). “Pragmatic Presuppositions,” in Stalnaker (1999).

———. (1978). “Assertion,” in Stalnaker (1999).

———. (1987). *Inquiry*. Cambridge: MIT Press.

———. (1998). “On the Representation of Context,” in Stalnaker (1999).

———. (1999). *Context and Content*. Oxford: Oxford University Press.

Stebbins, Sarah. (1992). “A Minimal Theory of Truth,” *Philosophical Studies* 66: 109-137.

Stenius, Erik. (1972). *Critical Essays*. Amsterdam: North-Holland Publishing Company.

Strawson, P. F. (1950). “On Referring,” *Mind* 59: 320-344.

Stueber, Karsten. (2002). “The Problem of Self-Knowledge,” *Erkenntnis* 56: 269-296.

Swan, Kyle. (2002). “Emotivism and Deflationary Truth,” *Pacific Philosophical Quarterly* 83: 270-281.

Sweet, Albert. (1999). “Local Semantic Closure,” *Linguistics and Philosophy* 22: 509-528.

Szabo, Zoltan Gendler. (1999). “Expressions and their Representations,” *The Philosophical Quarterly* 49: 145-163.

Tappenden, Jamie. (1993). “The Liar and Sorites Paradoxes: Toward a Unified Treatment,” *The Journal of Philosophy* 90: 551-577.

———. (1994). Review of McGee (1991), *The Philosophical Review* 103: 142-144.

———. (1999). “Negation, Denial, and Language Change in Philosophical Logic,” in *What is Negation?*, D. Gabbay and H. Wansing (eds.), Dordrecht: Kluwer.

Tarski, Alfred. (1933). “The Concept of Truth in Formalized Languages,” in *Logic, Semantics, Meta-Mathematics*, J. H. Woodger (tr.) and J.Corcoran (eds.), Indianapolis: Hackett Publishing Company, 1983.

———. (1944). "The Semantic Conception of Truth," *Philosophy and Phenomenological Research* 4: 341-376

Tennant, Neil. (1982). "Proof and Paradox," *Dialectica* 36: 265-296.

———. (1995). "On Paradox without Self-Reference," *Analysis* 55: 199-207.

Thomason, Richmond H. (1976). "Necessity, Quotation, and Truth: An Indexical Theory," in *Language in Focus*, A. Kasher (ed.), Holland: D. Reidel.

Thomson, J. F. (1962). "On Some Paradoxes," in *Analytic Philosophy*, Butler, R. J. (ed.), New York: Barnes and Noble.

Tienson, John. (1987). "An Argument Concerning Quantification and Propositional Attitudes," *Philosophical Studies* 51: 145-168.

Toms, Eric. (1956). "The Liar-Paradox," *Philosophical Review* 65: 542-547.

Truncellito, David. (2000). "Which Type is Tokened by a Token of a Word-Type?" *Philosophical Studies* 97: 251-266.

Unger, Peter. (1975). *Ignorance: A Case for Skepticism*. Oxford: Oxford University Press.

Ushenko, A. P. (1937). "A New 'Epimenides'," *Mind* 46: 549-550.

Urquhart, Alisdair. (2001). "Basic Many-Valued Logic," in *Handbook of Philosophical Logic*, vol. 2, D. Gabbay and F. Guenthner (eds.), Dordrecht: Kluwer Academic Publishers.

van Benthem, J. F. A. K. (1978). "Four Paradoxes," *Journal of Philosophical Lo*gic 7: 49-72.

van Fraassen, Bas C. (1966). "Singular Terms, Truth-Value Gaps, and Free Logic," *The Journal of Philosophy* 63: 481-495.

———. (1968). "Presupposition, Implication, and Self-Reference," *The Journal of Philosophy* 65: 136-152.

———. (1970a). "Inference and Self-Reference," *Synthese* 21: 425-438.

———. (1970b). "Truth and Paradoxical Consequences," in Martin (1970).

Visser, Albert. (2005). "Semantics and the Liar Paradox," in *Handbook of Philosophical Logic*, vol. 11, D. Gabbay and F. Guenthner (eds.), Dordrecht: Kluwer Academic Publishers.

Wedgwood, Ralph. (1997). "Non-Cognitivism, Truth, and Logic," *Philosophical Studies* 86: 73-91.

Weir, Alan. (1996). "Ultramaximalist Minimalism!" *Analysis* 56: 10-22.

———. (2000). "Token Relativism and the Liar," *Analysis* 60: 156-170.

Whiteley, C. H. (1958). "Let Epimenides Lie!" *Analysis* 19: 23-24.

Williams, Michael. (1986). "Do We (Epistemologists) Need a Theory of Truth?" *Philosophical Topics* 14: 223-242.

———. (1996). *Unnatural Doubts*. Princeton: Princeton University Press.

———. (1999). "Meaning and Deflationary Truth," *The Journal of Philosophy* 96: 545-564.

———. (2002). "On Some Critics of Deflationism," in Schantz (2002).

Williamson, Timothy. (1994). *Vagueness*. London: Routledge.

———. (1996). "Knowing and Asserting." *Philosophical Review* 105, 489-523.

———. (1997). "Imagination, Stipulation, and Vagueness," in *Philosophical Issues 8: Truth*, E. Villanueva (ed.), Atascadero, CA: Ridgeview Publishing.

———. (2000a). *Knowledge and Its Limits*. Oxford: Oxford University Press.

———. (2000b). "Semantic Paradox and Semantic Change," in *Proceedings of the Twentieth World Congress of Philosophy, vol. 6*, A. Kanamori (ed.), Bowling Green: Philosophy Documentation Center.

Wormell, C. P. (1958). "On the Paradoxes of Self-Reference," *Mind* 67: 267-271.

Wright, Crispin. (1992). *Truth and Objectivity*. Cambridge: Harvard University Press.

———. (1998). "Comrades against Quietism: Reply to Simon Blackburn on *Truth and Objectivity*," *Mind* 107:183-203.

Yablo, Stephen. (1982). "Grounding, Dependence, and Paradox," *Journal of Philosophical Logic* 11: 117-137.

———. (1985). "Truth and Reflection," *Journal of Philosophical Logic* 14: 297-349.

———. (1989). "Truth, Definite Truth, and Paradox," *The Journal of Philosophy* 86: 539-541.

———. (1993a). "Definitions, Consistent and Inconsistent," *Philosophical Studies* 72: 147-175.

———. (1993b). "Hop, Skip and Jump: The Agnostic Conception of Truth," *Philosophical Perspectives* 7: 371-396.

———. (2003). "New Grounds for Naïve Truth Theory," in Beall (2003)

Yaqūb, Aladdin M. (1993). *The Liar Speaks the Truth.* Oxford: Oxford University Press.

Yi, Byeong-Uk. (1999). "Descending Chains and the Contextualist Approach to Semantic Paradoxes," *Notre Dame Journal of Formal Logic* 40: 554-567.