TRUTH, SEMANTIC CLOSURE, AND CONDITIONALS

by

Shawn Standefer

B.A., Stanford, 2006

M.A., Stanford, 2006

Submitted to the Graduate Faculty of the

Kenneth P. Dietrich School of Arts and Sciences

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Shawn Standefer

It was defended on

August 20, 2013

and approved by

Anil Gupta, Department of Philosophy, University of Pittsburgh

Nuel Belnap, Department of Philosophy, University of Pittsburgh

Robert Brandom, Department of Philosophy, University of Pittsburgh

James Shaw, Department of Philosophy, University of Pittsburgh

Jeremy Avigad, Department of Philosophy, Carnegie Mellon University

Dissertation Director: Anil Gupta, Department of Philosophy, University of Pittsburgh

TRUTH, SEMANTIC CLOSURE, AND CONDITIONALS

Shawn Standefer, PhD

University of Pittsburgh, 2013

Almost all theories of truth place limits on the expressive power of languages containing truth predicates. Such theories have been criticized as inadequate on the grounds that these limitations are illegitimate. These criticisms set up several requirements on theories of truth. My initial focus is on the criticisms and why their requirements should be accepted.

I argue that an adequate theory of truth should validate intuitive arguments involving truth and respect intuitive evaluations of the semantic statuses of sentences. From this starting point, I analyze the arguments in favor of several common requirements on theories of truth and formulate some minimal requirements on theories of truth. One is a logic neutrality requirement that says that a theory must be compatible with a range of logical resources, such as different negations. Another is the requirement that the theory validate certain laws governing truth, such as the T-sentences. These two requirements rule out many theories of truth. The main problem is that many theories lack an adequate conditional, the addition of which is, in fact, precluded by those theories.

I argue that the revision theory of truth can satisfy my criteria when augmented with a pair of conditionals, which are defined using a modification of the framework of circular definitions of the revision theory. I distinguish two roles for conditionals in theories of truth and argue that the conditionals of the proposed theory fill those roles well. The conditionals are interdefinable with a modal operator. I prove a completeness theorem for the calculus C_0 of *The Revision Theory of Truth* modified with rules for this operator. I examine the modal logic of this operator and prove a Solovay-type completeness theorem linking the modal logic and a certain class of circular definitions.

I conclude by examining Field's recent theory of truth with its new conditional. I argue that Field's theory does not meet my requirements and that it fails to vindicate some of Field's own philosophical views. I close by proposing a framework for studying Field's conditional apart from his canonical models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 THE PROBLEM OF SEMANTIC CLOSURE

1.1 INTRODUCTION

Truth seems to be a simple concept, one that is governed by principles of the form,

$$\ulcorner A \urcorner \text{ is true if and only if } A$$

where $\ulcorner A \urcorner$ is a name of the sentence A. These principles have come to be called *T-sentences*, or *Tarski biconditionals*. The collection of T-sentences for a language, together with classical logic and some modest syntactic resources results in triviality, so the set of naively interpreted T-sentences alone cannot be an adequate account of truth. There is a need to develop a theory of truth, an account of the logical and semantic features of truth.

Truth easily gives rise to paradoxes, such as the liar paradox. A liar sentence is a sentence that says of itself that it is not true. Assuming that it is true, we can infer from the T-sentence that it is not true. Assuming that it is not true, we can infer from the T-sentence that it is true, after all. This paradox, and others, create challenges for theories of truth, to which there are many ways to respond. There are three broad approaches on which I will focus: the Tarskian approach, the fixed-point approach, and the revision theory approach. These approaches provide different philosophical and formal pictures of truth. I will defer the details of these approaches until §1.2, but I will note that while the three approaches agree on many things, their disagreements are large.

There are many theories competing as accounts of truth. There are many questions to ask of these theories. How do we choose between them? Are any of them the correct account of truth? Are they even trying to answer the same questions about truth? Despite the wide variety of formal and philosophical approaches to truth, all theories face a common objection:

they are not compatible with languages containing certain resources. For example, in one theory, a standard fixed-point theory, it is natural to say that the liar is neither true nor false, but the theory does not work with languages that contain the semantic resources to say that truthfully.[1] That fixed-point theory exhibits a failure of semantic closure; there are apparently relevant semantic concepts that cannot be added to the theory. This failure is cited as a reason not to accept the fixed-point theory.[2]

Philosophers have objected to every major theory of truth for failing to satisfy some closure condition that is claimed as important. Such failure is cited as grounds for rejecting the theory as inadequate. This inference, from failure to satisfy a closure condition to rejection, is common but it is not obviously correct. It is not clear why this sort of argument should be accepted across the board.

I will investigate the *broad problem of Semantic Closure*.[3] The problem concerns the sorts of closure conditions with which an adequate theory of truth must be compatible. There are many issues that fall within the bounds of the broad problem, issues which keep coming up in evaluations of theories and which call for careful discussion and distinctions. For example, one issue is whether a theory can be adequate if metalanguages for that theory are distinct from their object languages. Another issue is whether a theory can be adequate if key principles governing truth are not statable in the object languages of the theory. Both of these examples point to possible failures of different kinds of closure in languages for a theory of truth. Analogs of these examples show up in objections to theories in recent work on truth.

The broad problem of Semantic Closure can be motivated by reflecting on semantics. Many proponents of Semantic Closure see a close connection between semantics and theories of truth. The following is a natural line of thought about semantics.[4] Our language includes

[1] I will explain what this means in §1.2.

[2] See Field (2006a, 73), for example. Field thinks that the fixed-point theory is basically correct, but this issue is a sticking point for him for accepting Kripke's version.

[3] I capitalize "Semantic Closure" to distinguish it from narrow problems of semantic closure, which I talk about in the next chapter.

[4] This line of thought is based on the following quotation from Martin (1997, 417-418).

> [Gupta and Belnap] dismiss the goal of trying to understand truth for a language entirely from within the language. ... The problem that the semantic paradoxes pose is not primarily the problem of understanding the notion of truth in expressively rich languages, it

a truth predicate, so developing an adequate semantic theory for our language will require some theory of truth. This theory must be developed in our language, since we have no other in which to develop it. Our theory of truth, indeed our semantic theory, must be about our language and must be stated in that same language. The theory of truth must cover the resources needed for its own semantics. The preceding line of thought may be a natural one for many, but I will argue that it makes important errors, which I will bring out in chapter 2.

Since Tarski's work on truth, philosophers have objected to theories of truth based on their use of metalanguages and hierarchies of semantic predicates. More recently related problems, such as failure to express certain semantic concepts, have also been offered as objections. These problems are frequently run together. They fall under the heading of Semantic Closure.

The broad problem of Semantic Closure has many parts. I will clarify the problem of Semantic Closure by distinguishing six aspects of the problem that have shown up in discussions of Semantic Closure and arguing that concerns over Semantic Closure depend on two particular problems. I will argue that the best of these arguments for taking these problems as demands on theories of truth are not sound, once we have made some distinctions in foundational concepts for theories of truth. In the course of the analysis, I will examine other aspects of Semantic Closure and argue that some of them provide conditions of adequacy on theories of truth that should be accepted. Following my analysis of the problem of Semantic Closure, I will examine two theories of truth, Field's theory and a modified form of the revision theory, and compare them on criteria that emerge from the discussion of Semantic Closure.

The outline of my dissertation is as follows. In chapter 1, I present the relevant background and begin my analysis. In §1.2, I briefly present both the formal details and philosophical motivations of three approaches to truth, the Tarskian, fixed-point and revision theory approaches. In §1.3, I present a survey of the literature on theories of truth to point out appeals to Semantic Closure that have been made. In §1.4, I present my categorization

is the problem of understanding our notion of truth. And we have no language beyond our own in which to discuss this problem and in which to formulate our answers.

of the six subproblems of Semantic Closure. In §1.5, I distinguish different philosophical projects for which a theory of truth can be used. Distinguishing these projects is important for getting clear on the import of the subproblems from §1.4. In §1.6, I present some background details on types of languages and views about semantics. With all that background in place, I start my analysis of the problem of Semantic Closure, beginning with two subproblems in this chapter. In §1.7, I will examine the role of semantic hierarchies in theories of truth and argue that the existence of semantic hierarchies is not central to the problem of Semantic Closure. In §1.8, I will present one of the subproblems of Semantic Closure, the problem of semantic self-sufficiency, and Gupta and Belnap's criticisms of it.

In chapter 2, I will continue my analysis of the broad problem of Semantic Closure. There I will argue against many claims, but I reach some positive conclusions as well. One of these is a logic neutrality thesis, which I explain in §2.1. Another positive thesis is that the T-sentences and semantical laws must turn out valid in an adequate theory of truth, which highlights the importance of conditionals in theories of truth.

From the analysis of the first two chapters, the role of conditionals in theories of truth emerges as important for assessing the adequacy of a theory. In the final two chapters, I turn to a more focused study of conditionals. In chapter 3 I motivate an addition of two new conditionals to the revision theory and argue that the resulting theory comes out well with respect to the criteria of the first two chapters. In chapter 4, I present the proofs of the main results concerning the expanded revision theory. In chapter 5, I investigate the modal logic motivated by the definitions of chapter 3 and prove a completeness result relating the modal logic to a class of circular definitions. In chapter 6, I present and criticize Field's theory of truth, with a special focus on the conditional he uses to augment the strong Kleene fixed-point theory.

Now, I will begin presenting the necessary formal and philosophical background for the first two chapters, beginning with the details of three approaches truth.

1.2 THREE APPROACHES TO TRUTH

The material conditional T-sentences and classical logic are inconsistent when there is a liar sentence in the language. To see this, take a liar sentence, L, for which it is provable, or true, that $L \equiv \sim T(\ulcorner L \urcorner)$, where T is a truth predicate.[5] The T-sentence for a liar is $T(\ulcorner L \urcorner) \equiv L$. The T-sentence together with the previous biconditional gives

$$T(\ulcorner L \urcorner) \equiv \sim T(\ulcorner L \urcorner),$$

which is a contradiction. There are different ways of developing a theory of truth to allow a consistent, or non-trivial, truth predicate in an object language. In this section, I will present details on three approaches, the Tarskian approach, the fixed-point approach, and the revision theory approach.

The constructions I will describe share some common features. First, there is a ground model M that interprets a *ground language* $\mathscr{L}$. A model M is a pair $\langle D, I \rangle$ of a domain D and an interpretation function I. The language is extended with a truth predicate. Depending on the syntactic devices in $\mathscr{L}$, names for sentences of the extended language may be added and the domain may be enriched with the sentences of the expanded language. For the exposition, I will assume that the names of sentences are interpreted as the sentences themselves, so that the extensions of the truth predicates are sets of sentences, rather than, say, sets of Gödel numbers.[6] The extension of the truth predicate is then determined in accordance with the chosen approach.

Tarski proposed one way of developing a consistent theory of truth. He proposed restricting the set of T-sentences. The extension of a truth predicate on the Tarskian approach contains only sentences not using that truth predicate.[7] To develop a theory of truth for

[5] I will use '$\ulcorner\ \urcorner$' as a generic way of forming names of sentences. I will also use quotation names at points where a particularly simple syntactic theory is assumed. I will use '$\supset$' and '$\equiv$' and for the material conditional and biconditional, respectively. In chapter 2, especially §2.5, I will use '$\Leftrightarrow$' as a generic biconditional. In chapters 3-6, I will use '$\leftrightarrow$' and its conditional parts for specific conditionals, which I will explain in those chapters.

[6] This is done in Kremer (1988), for example.

[7] More carefully, the extension contains the denotation of names, or other designators, of sentences that do not use the truth predicate. If the syntactic theory uses Gödel numbers of sentences as their names, the extension of the truth predicate could be a subset of ω, rather than a subset of the language.

a language with a truth predicate, a new truth predicate, whose extension contains only sentences not using this new predicate, has to be introduced. This gives rise to a hierarchy of truth predicates, each applying correctly only to sentences strictly lower in the hierarchy.

The technical picture begins, as above, with a ground model M that interprets a language $\mathscr{L}$. $\mathscr{L}$ is extended with a truth predicate, T_0, to $\mathscr{L}^+$. The extension of the truth predicate is the set $h_0 = \{A \in Sent(\mathscr{L}) : M \models A\}$, the set of sentences satisfied in the ground model.[8] This gives rise to a model, $M+h_0$, where h_0 is used to interpret T_0. A second truth predicate, T_1, can be added to extend the language to $\mathscr{L}^{++}$. The extension of this new predicate is given by $h_1 = \{A \in Sent(\mathscr{L}^+) : M + h_0 \models A\}$. Further truth predicates can be added to languages in this sequence, and their extensions will be defined on the languages earlier in the sequence.[9]

The Tarskian approach was, for a long time, the dominant approach to truth. Kripke criticized the Tarskian theory of truth on the grounds that its restrictions were too costly. Self-referential uses of truth are common enough in natural language argumentation, and they are not always vicious. Even non-vicious instances of self-reference are difficult to accommodate in the Tarskian theory. Kripke supplied an example of truth attributions for which the Tarskian theory of truth has great difficulty accounting. This is the *Nixon-Dean case*.[10] In a deposition about Watergate, Nixon says that everything Dean said is true while Dean says that everything Nixon said was not true, and neither said anything else. The problem for the Tarskian comes when she tries to assign appropriate levels of the hierarchy to the truth predicates in the attributions. Suppose Nixon's is at level n. Then Dean's must be at a strictly higher level, but this requires that Nixon's be at a yet higher level. The most natural way to formalize the truth attributions is with a self-referential truth predicate, which can apply correctly to sentences containing that very predicate.

In the mid 70s, Kripke and Martin and Woodruff showed how to define languages that contain self-referential truth predicates.[11] They developed the *fixed-point theory of truth*,

[8] The notation '$M \models A$' is read as 'A is true in M'. In chapter 3, this notation will be repurposed to stand for validity.

[9] There are more intricate ways of defining the Tarskian truth predicates. See Glanzberg (2004), for example. One can also proceed axiomatically. See Burge (1979) or Horsten (2011) for details.

[10] Kripke (1975, 695-697)

[11] See Kripke (1975) and Martin and Woodruff (1975).

so-called because of the technique used to define its models. They showed how to start from a ground language obeying certain conditions and then how to add a truth predicate that is interpreted in a non-classical way.[12] The intuitive idea behind Kripke 's construction is as follows.[13] Initially, the truth predicate has an empty extension and anti-extension.[14] The extension is built up in stages. If A is true in the ground model, then it is added to the extension of the truth predicate in the first stage, and, similarly, if A is false in the ground model, then it is added to the anti-extension. The process proceeds by adding sentences to the extension and anti-extension depending on whether they are true, or not, in the ground model with the current interpretation of the truth predicate. If A is in the extension of T at one stage, then $T(\ulcorner A \urcorner)$ will be in the extension in the next stage, and similarly with the anti-extension. Once a sentence is put into the extension, then it remains there for the remainder of the construction. Some sentences, such as the liar, are placed into neither the extension nor the anti-extension. Eventually the construction hits a stage after which no more sentences are added to either the extension or the anti-extension. This is the fixed-point.

More formally, we begin with a language $\mathscr{L}$ interpreted by a classical ground model M. A truth predicate, T, is added to $\mathscr{L}$ to extend it to $\mathscr{L}^+$. Unlike the Tarskian approach, here T is not interpreted classically. It is, rather, given a partial interpretation via a pair, $(\mathscr{T}, \mathscr{F})$, where $\mathscr{T}$ is the extension of the truth predicate and $\mathscr{F}$ is the anti-extension. The construction of the fixed-point proceeds in stages from an initial hypothesis (Tr_0, Fa_0) for the extension and anti-extension, such as the empty set for both.[15]

- At stage 0, let $Tr_0 = Fa_0 = \emptyset$.
- At successor stages $\alpha+1$, $(Tr_{\alpha+1}, Fa_{\alpha+1}) = \Phi(Tr_\alpha, Fa_\alpha)$, where Φ is the *jump operation*. The jump operation Φ is defined such that $\Phi(Tr_\alpha, Fa_\alpha)$ is the pair consisting of the set

[12] The ground language cannot contain non-monotonic operators and neither can the extended language. An operator O is non-monotonic if there are semantic values $\mathbf{a}_0, \ldots, \mathbf{a}_n, \mathbf{b}_0, \ldots, \mathbf{b}_n$ such that $\mathbf{a}_i \leq \mathbf{b}_i$ but $O(\mathbf{a}_0, \ldots, \mathbf{a}_n) \not\leq O(\mathbf{b}_0, \ldots, \mathbf{b}_n)$. This is definition depends on the ordering of semantic values.

[13] Martin and Woodruff use Zorn's lemma, rather than a stage by stage construction, to obtain their fixed-points.

[14] The extension and anti-extension are, respectively, the sets of objects a predicate is true of and those it is not true of.

[15] Other initial hypotheses can be used, although there are some restrictions that prevent the use of arbitrary initial hypotheses.

of sentences that are true in $M + (Tr_\alpha, Fa_\alpha)$ and the set of sentences that are false in $M + (Tr_\alpha, Fa_\alpha)$. More formally, $\Phi(Tr_\alpha, Fa_\alpha)$ is the following pair.[16]

$$(\{A \in Sent(\mathscr{L}^+) : M + (Tr_\alpha, Fa_\alpha) \models A\}, \{A \in Sent(\mathscr{L}^+) : M + (Tr_\alpha, Fa_\alpha) \models \neg A\})$$

- At limit stages λ, let

$$Tr_\lambda = \bigcup_{\eta<\lambda} Tr_\eta$$

and

$$Fa_\lambda = \bigcup_{\eta<\lambda} Fa_\eta.$$

Eventually this constructions reaches a *fixed-point*, which is a stage α such that $\Phi(Tr_\alpha, Fa_\alpha) = (Tr_\alpha, Fa_\alpha)$. Supposing that γ is the first stage at which (Tr_γ, Fa_γ) is a fixed-point of Φ, let $(\mathscr{T}, \mathscr{F}) = (Tr_\gamma, Fa_\gamma)$.[17]

Since the truth predicate is given only a partial interpretation, a classical interpretation of the connectives is not always available. The three-valued interpretation of the connectives on which I will focus is that of *strong Kleene* logic, whose connectives are given by the following tables.

$\sim$	
t	**f**
n	**n**
f	**t**

&	**t**	**n**	**f**
t	**t**	**n**	**f**
n	**n**	**n**	**f**
f	**f**	**f**	**f**

$\supset$	**t**	**n**	**f**
t	**t**	**n**	**f**
n	**t**	**n**	**n**
f	**t**	**t**	**t**

The universal quantifier is a generalized conjunction. Disjunction and the existential quantifier are definable from conjunction and the universal quantifier, respectively, and negation.

While I will mainly discuss the strong Kleene fixed-point theory of truth, there are two other three-valued schemes that will come up. One is weak Kleene. The truth tables for the connectives in weak Kleene have **n** wherever one of the arguments is **n**. The other scheme is LP, which uses the same truth-tables as strong Kleene but differs in designated

[16] This definition does not require that $T(c)$ be put in the anti-extension, if 'c' names a non-sentence. The definition can be adjusted to do so. Kremer (1988, 235, 250) notes Kripke requires the truth of non-sentences to be in the anti-extension, although this requirement results in a non-compact consequence relation.

[17] The presentation here is in terms of sets. There is a natural algebraic presentation that uses truth values, the notation for which I will use later. In the fixed-point, the sentences in the extension of the truth predicate receive the value **t**, those in the anti-extension **f**, and those in neither set **n**.

values. Strong Kleene logic has one designated value, **t**, but LP has two, **t** and **n**, which, in discussions of LP is often referred to as **b** for "both true and false."

One of the main philosophical motivations for theories that contain their own truth predicates is the reconstruction and analysis of natural language arguments that involve self-referential attributions of truth. The analysis of such arguments has become a measure of whether a theory of truth has succeeded in reconstructing the colloquial notion of truth.

The fixed-point theories have many virtues. They assign the same values to A and $T(\ulcorner A \urcorner)$. They capture intuitions about some self-referential sentences failing to be grounded in non-semantic truths. They can be used to distinguish many classes of self-referential sentences that are all forbidden in the standard Tarskian account. A truth-teller, which says of itself that it is true, is intuitively semantically different from a liar, and this distinction can be made with the fixed-point theory.

Fixed-point theories also have some important faults. They require restricting the logical resources of the ground language. They also require, in many cases, changing the logic of the language. For example, if the ground language is classical, the language with truth will be interpreted in a 3-valued way. In addition, the fixed-point theories cannot validate all of the T-sentences.[18]

A third approach, the *revision theory of truth*, was developed by Gupta, Herzberger, and Belnap in response to some of the defects of the fixed-point theory.[19] The revision theory is a general theory of circular definitions of the form $Gx =_{Df} A(x, G)$.[20] A set of circular definitions $\mathscr{D}$ specifies a rule of revision $\delta_{\mathscr{D},M}$.[21] The rule of revision δ determines how a predicate's extension is to be revised given some hypothesis about what the extension of the predicate is, where a hypothesis assigns to each circularly defined predicate an extension. The revised hypothesis $\delta(h)$ will be the set of things that $A(x, G)$ is true of, using h to interpret G. The process of revision can be repeated transfinitely, and patterns in revision

[18] This is true of the standard fixed-point theories. As we will see, there are modifications one can make to the standard theory, adding new conditionals, that fix this defect.

[19] See Gupta (1982), Herzberger (1982), and Belnap (1982b), respectively. A comprehensive account of the revision theory of truth and a generalization to a general theory of circular definitions can be found in Gupta and Belnap (1993).

[20] The revision theory works with sets of such definitions, which can be of any arity and can be interlinked.

[21] I will drop the subscripts on δ when it is clear what set of definitions and what model are under consideration.

behavior of sentences emerge.

The revision theory of truth takes the set of T-sentences to be the circular definition of truth, so the T-sentences specify a rule of revision τ. In the revision process for truth, A will be in $\tau(h)$ if A is satisfied in the model $M + h$. Unlike the fixed-point approach, sentences may leave the extension of the truth predicate. In the revision process, some sentences will stabilize one way, others another way, and yet others, such as the liar, will not stabilize at all. Under some conditions, the revision process will reach a fixed-point, but in general it will not.

More technically, the revision process generates a sequence of models. As before, we start with a ground model M that interprets the ground language $\mathscr{L}$, which is extended to $\mathscr{L}^+$ with the addition of a truth predicate, T. T is interpreted by any hypothesis, h_0. A new extension for the truth predicate is generated by the revision rule τ, which sets the new extension to the set $\tau(h_0) = h_1 = \{A \in Sent(\mathscr{L}^+) : M + h_0 \models A\}$. More generally, the successor stage interpretations are given by

$$\tau(h_\alpha) = h_{\alpha+1} = \{A \in Sent(\mathscr{L}^+) : M + h_\alpha \models A\}.$$

At limit stages, whether a sentence is in the extension or not depends in part on the revision sequence leading up to the limit. Some sentences will stabilize in the sense that there is a stage α after which either they are in the extension until the limit or they are not in the extension until the limit. Stable sentences retain their stable position at limit stages. Unstable sentences can be assigned arbitrarily.[22]

Eventually the revision sequence reaches a stage after which only a distinguished set of hypotheses, the *cofinal hypotheses*, recur. There are different sequences one obtains by varying the initial hypothesis and limit rules. A hypothesis is *recurring* if it is a cofinal hypothesis in some revision sequence for τ from a ground model M. There will also be stages, *reflection stages*, that reflect the stabilities and instabilities of the process running through all the ordinals. The sentences that are stable, or unstable, at reflection stages will be stable, or unstable, respectively, in the revision process running through all the ordinals.

[22]There are various policies one can set for unstable sentences at limit stages, such as all unstable sentences are excluded from the extension at limits.

There are two different notions of stability that can be used. A sentence is *stable* if there is some stage after which it does not change its truth value. The second notion, *near stability*, permits some oscillation for finitely many revisions after a limit stage. These two notions can be used with recurring hypotheses to define different senses of validity. A sentence is valid in the first sense if it is true in each model $M + h$, where h is a recurring hypothesis. A sentence A is valid in the second sense if for all recurring hypotheses h, there is a number n such that for all $p \geq n$, A is true in $M + \tau^p(h)$. The sets of sentences valid in these two senses will be of interest later on, particularly in chapter 2.

Revision theories have many virtues. Unlike the fixed-point theories, the revision theories work with unrestricted logical resources. They do not need to change the logic of classical ground languages. They also have some faults, although the faults are distinct from those of the fixed-point theories. They can be technically complex and some are ω-inconsistent. It appears that adopting any of these three approaches will require giving up on some desirable quality in a theory of truth.

In addition to the theories sketched above, there are others. There is Priest's *dialetheic theory*, which permits true contradictions.[23] There are modified versions of some theories, such as Burge's and Glanzberg's contextualist versions of the Tarskian theory.[24] There are also theories that combine ideas from different theories, such as Beall's theory, which combines ideas from fixed-point and paraconsistent theories, and Field's, which combines ideas from fixed-point and revision theories.[25] There are yet other theories which I will not be able to discuss in detail.[26]

With that formal background in place, I will move to the overview of the use of Semantic Closure in the literature on theories of truth.

[23] See Priest (2006).

[24] See Burge (1979) and Glanzberg (2004), respectively.

[25] See Beall (2009) and Field (2008), respectively.

[26] For inconsistency theories, see, for example, Chihara (1979), Patterson (2007) and Scharp (2007). For the prosentential theory, see Grover et al. (1975) and Grover (1992). For a modern version of Bradwardine's theory, see Read (2009) and Read (2010). Recently much work has been done on axiomatic theories of truth and contraction-free theories. See Halbach (2011) for an overview of the former. For the latter, see, for example, Restall (1994), Zardini (2011) and Beall and Murzi (2013). Finally, for a so-called transitive-free theory, see Cobreros et al. (2013) and Ripley (2013).

1.3 APPEALS TO SEMANTIC CLOSURE

In this section I will present several objections to theories based on issues of Semantic Closure. The uses of Semantic Closure in objections tend to run together ideas I will later distinguish. I will present the objections in three groups. The first group consists of worries about the role of metalanguages in theories of truth. One of the places this worry appears is in discussions of the semantics of natural language, which will be the second group. The third group builds on the philosophical picture of the fixed-point theory found in Kripke (1975).

Tarski thought that the only way to maintain classical logic with a theory of truth in a rich language was to relegate the truth predicate for one language to a distinct metalanguage. Tarski's techniques, if iterated, produce a hierarchy of languages and truth predicates. Many philosophers have been dissatisfied with this hierarchy because it seems to require resources for which it cannot account.

Fitch criticized the Tarskian theory on the grounds that it could not account for the truth predicates it generated. He wanted a theory of truth that contained truth for the language of the theory so that "we are no longer forced to try to climb either an unending topless ladder of formal metalanguages, or a ladder of formal metalanguages that ends with a natural language at the top."[27] Fitch pointed out that philosophers frequently develop theories that talk about all theories and that the Tarskian approach to truth is incompatible with this practice, even in benign cases of self-reference. Fitch's criticisms focused on hierarchies of truth predicates and the role of metalanguages in the Tarskian view.

More recently, Kripke (1975) and Martin and Woodruff (1975) showed how to define languages that contain their own truth predicates. Their definitions showed that the truth predicate need not have its range of application restricted or to be relegated to a separate language, overcoming major faults of the Tarskian approach.

Kripke, in an influential passage, voiced a doubt about whether his theories solved the deep problems of the Tarskian theory.

> Now the languages of the [fixed-point] approach contain their own truth predicates and even

[27] Fitch (1964, 397-398)

> their own satisfaction predicates, and thus to this extent the hope [for a universal language, one in which everything that can be stated at all can be expressed,] has been realized. Nevertheless, the present approach does not claim to give a universal language... there are assertions we can make about the object language which we cannot make in the object language. ... The necessity to ascend to a metalanguage may be one of the weaknesses of the present theory. The ghost of the Tarski hierarchy is still with us.[28]

Kripke cites the need to "ascend to a metalanguage" in order to truthfully make certain assertions about his theories of truth as a problem for those theories. This locates the problem of Semantic Closure in the metalanguages, in particular the use of metalanguages in a capacity similar to that of the Tarskian theory. Indeed, the strong Kleene fixed-point theory does not say that the liar is neither true nor false, nor can it do so. The problem is that the theory of truth lacks closure under natural semantic categories, but the relation of this problem to the point about metalanguages is not obvious.

Priest has supplied many criticisms of philosophers who employ a metalanguage more expressive than the language of the theory of truth they offer. Priest sees the use of richer metalanguages as being of a kind both with the existence of semantic hierarchies and with the failure of certain closure conditions on the semantic vocabulary. All are rooted in expressive incompleteness, the inability of a consistent theory to express all the semantic concepts it needs. Priest provides arguments, based on metalanguage concerns and on failures of closure, against all consistent theories of truth.

In addition to criticisms of certain kinds of metalanguages, Priest has another line of objection to many theories. He, following Kripke, focuses on features of natural languages. Priest points to alleged features of natural languages as the grounds for rejecting certain theories of truth. He says, "A natural language (or a formal language that models that aspect of its behaviour) can give its own semantics."[29] Priest continues by saying that, since our language can be used to give its own semantic theory, any theory of truth that is incompatible with this will be deficient as an analysis of our concept of truth. Priest's idea is important to the broad problem of Semantic Closure. Many philosophers think that a theory of truth has to be capable of being a semantic theory for its own language. According to Priest, semantic hierarchies and certain uses of metalanguages by the fixed-point and

[28] Kripke (1975, 79-80)
[29] (Priest, 2006, 70)

Tarskian approaches are incompatible with the possibility of expressing a semantic theory for our language.

Priest has said that an adequate theory of truth must validate all the T-sentences, as well as the analogs for the satisfaction relation.[30] These sentences, together with classical logic and modest syntactic resources, generate contradictions. If a theory fails to validate the T-sentences, then, according to Priest, it fails to meet some minimal closure conditions for being a good semantic theory.

McGee has been one of the most vocal critics of theories of truth on the basis of their lack of Semantic Closure. He shares some of Priest's concerns over the use of metalanguages and the prospects of developing a semantic theory for the language of the theory itself. McGee views the problem of Semantic Closure as central because "to obtain an understanding of the notion of truth as it applies to English [one must] solve the problem of how to present the semantics of a language within the language itself."[31] He continues by saying, "[I]f we can only give the semantics of our simplified language within an essentially richer metalanguage, the fundamental and difficult problem of how to give the semantics for a language within the language itself will still remain before us."[32] McGee sees the role of a theory of truth in semantic theories for natural language as one of the fundamental concerns for adequate theories of truth. McGee uses the problem of Semantic Closure as an objection to all the extant major theories of truth and he has repeatedly used it in objections to the revision theory of truth.[33] I will examine his arguments in more detail in §1.8.3 and §2.4.3. Martin has joined McGee in criticizing the revision theory on the basis of alleged failures of Semantic Closure.[34] These arguments depend heavily on the role of metalanguages in theories of truth and a strong conception of how a theory of truth should feature in a semantic theory.

Gupta and Belnap responded to McGee's original criticisms, providing the only serious opposition to the use of one kind of Semantic Closure, semantic self-sufficiency, as a condition of adequacy on theories of truth.[35] They point out that the idea of semantic self-sufficiency

[30] See Priest (1984) for the argument. I will discuss this point more in §2.2.
[31] McGee (1991, 147)
[32] McGee (1991, 147)
[33] See McGee (1991) and McGee (1997).
[34] See Martin (1997).
[35] See Gupta and Belnap (1993, 256-259).

itself is not clear and argue against the main reasons for taking semantic self-sufficiency and other forms of Semantic Closure to be adequacy conditions on theories of truth. Despite their arguments and the apparent importance of Semantic Closure for recent work on theories of truth, there has not been much response to their arguments and no sustained investigation of the broad problem of Semantic Closure.

Gupta responded to McGee and Martin's more recent criticisms.[36] Gupta's responses focus on the need for clarification of the problem and undermine of the primary motivations for adopting some aspects of the broad problem of Semantic Closure as criteria of adequacy for theories of truth.

Many philosophers working on truth have followed Kripke in viewing the broad problem of Semantic Closure as important for the theory of truth, but they, like Kripke, have run together many issues that could be the source of the problem. For example, Maudlin says that if Kripke is correct that the need to ascend to a metalanguage has not been avoided, "then it is an understatement to say that a metalanguage is needed: a whole Tarskian infinite hierarchy of metalanguages will be needed. And it becomes less and less clear exactly what has been accomplished."[37] According to Maudlin, the broad problem of Semantic Closure undercuts the philosophical gains that the fixed-point theories were supposed to have over the Tarskian theory. Maudlin does not make clear why the hierarchies or the metalanguages are a problem.

Maudlin (2004), in a chapter entitled "A Language That Can Express Its Own Truth Theory," says, "*Since one of the primary questions facing the theory of truth is whether any language can serve as its own metalanguage*, it seems worthwhile to pause to ask whether the semantics of our expanded language can be expressed in the language itself."[38] Maudlin asserts that one of the primary questions for any theory of truth is whether the language of the theory can serve as its own metalanguage. This is a form of Semantic Closure similar to that discussed by McGee. Maudlin does not explain why this question is one of the primary ones for the theory of truth. The problem Maudlin is addressing could be one concerned with the role of metalanguages or the role of a theory of truth in a semantic theory.

[36] See Gupta (1997).
[37] Maudlin (2004, 52)
[38] Maudlin (2004, 79), emphasis added.

Field responds to the same aspect of Kripke's work as Maudlin. Field bills his theory of truth as one that "there is no reason to go beyond simply to say what we want to say about truth and related concepts."[39] Field says that he is attempting to construct languages that are "comprehensive enough to be 'self-sufficient' in that they do not contain within themselves the requirement that they be expanded."[40] Field cites issues surrounding metalanguages, hierarchies, and the expressibility of semantic categories as motivations for his approach. Issues of Semantic Closure provide a motivation for Field's theory of truth while remaining in the background, with no elaboration on why they were taken as guiding principles.

Field's theory is comprehensive in the sense that it can express an appropriate semantic classification for each sentence of his object languages. The alleged need for ways to classify sentences has become prominent enough that Beall has recently highlighted semantic classification as a dominant trend in recent work on theories of truth. The project he highlights is to show "how, if at all, we can truly characterize –specify the 'semantic status' of– all sentences of our language (in our language)."[41] Beall thinks that we can so specify the statuses of our sentences, and he endorses the possibility of such a specification as a requirement on theories of truth. Beall does not say much about why he takes the question as important to the evaluation of theories of truth.

This brief look at recent work on truth indicates the ways in which the broad problem of Semantic Closure arises. It also indicates the ways in which issues run together. Major figures move freely between issues surrounding hierarchies and metalanguages and the role of truth in semantic theories. Field and Beall make connections to issues of classification, which were arguably in Kripke's work as well. In the next section, I will distinguish six subproblems in the broad problem of Semantic Closure. These distinctions will help isolate both what is at stake and what the relations are between the various issues canvassed in this section.

[39] Field (2008, 222)
[40] Field (2008, 18)
[41] Beall (2009, 66). I will discuss this project in §2.3.3.

1.4 PROBLEMS

Based on the quotations from the previous section, one might think that the broad problem of Semantic Closure is any of a number of things, such as a problem with hierarchies of semantic concepts or a problem accommodating a universality property.

I will provide an initial characterization of the six parts of the broad problem of Semantic Closure. The *hierarchy problem* is the problem of developing a theory of truth that does not contain any hierarchies of semantic concepts. I will discuss this more in §1.7. The *problem of self-sufficiency* is the problem of developing a theory of truth for our language that can impart an understanding of that language. I will discuss this more in §1.8. The *universality problem* is the problem of developing a theory of truth that accommodates a universality property, which exhibits a sense in which a language is unbounded, that natural language is alleged to have. I will explain this in §2.1. The *narrow closure problem* is the problem of developing a theory of truth that is closed under combinations of its semantic vocabulary and the semantic vocabulary of its designated metalanguages. This will be covered in §2.2. The *classification problem* is the problem of developing a theory of truth that can classify each of the sentences of the language of the theory into some semantic category, once the non-semantic facts have been specified. This will be the focus of §2.3. The *metalanguage problem* is the problem of developing a theory of truth that has a metalanguage identical to the language of the theory of truth. I will discuss this in §2.4.

The broad problem of Semantic Closure appears to be a combination of these six problems. We must separate out these problems and the motivating ideas behind them. The analysis I will develop is that the core issues of Semantic Closure are the classification and metalanguage problems. This analysis makes sense of the debates in §1.3. Many of the arguments for the centrality or importance of the other problems depends on arguments for these two problems.

My analysis will take up the rest of the first two chapters, so I will provide an outline of it here. There are compelling reasons against taking the broad problem of Semantic Closure to be primarily concerned with the hierarchy problem and the problem of semantic self-sufficiency, reasons which I will present in §1.7 and §1.8, respectively.

Following this, I will distinguish two forms of universality and argue that one, but not the other, puts some constraints on theories of truth (§2.1). Then, I will examine the narrow closure problem and argue that the theory of truth does need to satisfy some closure properties (§2.2). Next, I will present and criticize the arguments in favor of imposing the dual requirement of solving the classification and metalanguage problems on theories of truth in §2.3 and §2.4, respectively. In so doing, I will make clearer why this is a good analysis of the problem of Semantic Closure, to clarify the content of the dual requirement, and to show that there are no strong arguments for imposing either of these requirements on theories of truth. One theme that emerges from my discussion, is that a theory of truth should validate certain laws for semantic vocabulary, which I will discuss in §2.5.

1.5 PROJECTS

There are many projects that could be undertaken in developing a theory of truth.[42] The primary distinction needed for my arguments is the distinction between *normative* and *descriptive* projects.[43]

Normative: A theory of truth aims to provide replacement concepts that are in some way better than what they are replacing

Descriptive: A theory of truth focuses on clarifying and regimenting the everyday notions of truth and related concepts

The normative project attempts to replace or revise the concept of truth. A common reason adopting the normative project is the thought that the everyday concept of truth needs to be replaced is that the concept is inconsistent. Quine, for example, thinks this.[44] One could also engage in the normative project by proposing a new concept of truth for use in formal

[42] A brief note on terminology. In what follows, I attribute goals or projects to a theory of truth. This is a shorthand way of attributing those goals or projects to proponents of that theory of truth. Theories of truth may have built in goals or projects when they are considered in a broad sense that includes philosophical commitments not included as part of the formal theory.

[43] This distinction was described in Gupta and Belnap (1993).

[44] Quine (1976)

languages because it is a useful concept to have. Feferman, Quine and Field can each be viewed as engaging in normative projects.

The descriptive project for truth takes as its goal the reconstruction and formalization of everyday notions of truth. The target phenomena are inferences and arguments involving the truth predicate, and possibly other semantic vocabulary. The phenomena may also include broader features about meaning in natural language to which truth is tied. Gupta, Belnap, Priest, and McGee are each engaged in the descriptive project.

There are at least two broad approaches to the descriptive project: logic-based and principle-based. The two overlap but their orientations are different. The logic-based approach focuses on logic and arguments, and it requires that an adequate theory of truth provide correct verdicts on a large range of arguments involving the truth predicate. This includes intuitive judgments about the status of sentences. The principle-based approach requires that an adequate theory of truth preserve certain principles about truth and its relation to semantics in natural language. It also requires that a theory of truth be compatible with general principles about natural language, such as the principle that natural languages are able to refer to their own words. I will take the logic-based approach as more important. I will, however, appeal to the principle-based approach when the principles in question are sufficiently well established.

The logic-based approach has two main advantages over the principle-based approach. The first is that it focuses on arguments, which are relatively concrete. There are a wide range of arguments involving truth that can be used to compare and evaluate theories. This allows a fine-grained comparison of theories.

The second advantage is epistemological. We have a clearer grasp of arguments involving truth, which are good and which not, than of the principles of the principle-based approaches. Many of the principles are in need of philosophical clarification, a point I will illustrate in this chapter and the next. Because we have a clearer grasp of arguments involving truth, there can be a finer-grained comparison and evaluation of theories. This is not to say that all arguments will be evaluated as clearly valid or clearly invalid. There will be borderline cases and there will be cases that presuppose some background logical inference which may be under scrutiny. Nonetheless, the range of arguments provides grounds for fine-grained

comparison of theories on the basis of the many uncontroversial arguments. Philosophical discussion can then focus on the controversial points.

The principle-based approach does not seem to provide a degree of precision in comparing theories comparable to that of the logic-based approach. Many of the principles themselves require philosophical work before they can clearly render verdicts on theories of truth and the verdicts are often of the simple "yes/no" variety. For these reasons, the logic-based approach is the one that will underlie my analysis.

Both of approaches to the descriptive project are adopted in the literature. McGee and Martin adopt the principle-based approach. They take truth to be a foundational semantic concept. They think that a theory of truth needs to be a part of the semantic theory for the language of the semantic theory, a requirement they think stems from natural languages.[45] This requirement is an example of a global principle. Gupta, by contrast, adopts the logic-based approach. He thinks that a theory of truth should primarily be judged according to whether it correctly classifies as invalid or invalid arguments using the truth predicate.[46] McGee and Martin take the global principles to be a filter on admissible theories of truth while Gupta takes logical features to be a filter on admissible theories of truth.[47] One thing that will help decide between these approaches is an assessment of the arguments in favor of the principles adopted by McGee and Martin, assessments I will offer in chapter 2.

A theory of truth accomplishes the descriptive project only if it accurately models the colloquial notion of truth. The point can be put in terms of real truth predicates and truth-like predicates.[48] A theory of truth that fails to provide a truth predicate that is sufficiently similar to the colloquial truth predicate, one that does not model some important aspect of the logical behavior of the truth predicate, is not a theory of the colloquial notion of truth. The theory is a theory of a truth-like concept rather than a theory of truth.

Insofar as a theory purports to describe the colloquial notion of a truth, that theory loses some of its philosophical significance when it is shown to be a theory of a truth-like notion

[45] See McGee (1997) and Martin (1997), respectively.

[46] See Gupta (1997).

[47] There can be some overlap between the principles and the requirements of the logic-based approach. For example, validating the T-sentences could fall under either.

[48] My use of these terms is somewhat contentious, but I think that the philosophers I discuss would accept these terms.

rather than a theory of truth. This is not to say that theories of truth-like notions could not be formally or philosophically interesting. They can be, but they will need defense and interpretation to demonstrate their philosophical interest.

The strongest arguments that a theory of truth should satisfy various semantic closure conditions depend on citing features, or alleged features, of natural language. The descriptive project provides good reason for these features to be relevant to a theory of truth. When a normative theory of truth is under discussion, it is less clear what features of natural language should be relevant and why. A natural response is that one should adopt the view that the theory of truth should respect features of natural language as far as possible while replacing truth and associated notions with those of the theory. This response selectively rejects some commitments of the descriptive approach, and this response needs an elaboration and defense of the relevant features. I am unaware of anyone engaged in the normative project that has offered such an elaboration.[49] I think the preceding supplies good reason to focus on descriptive theories of truth rather than normative ones.

There is a distinction that is important in light of the distinction between the descriptive and normative projects; that is the distinction between possible and actual closure. *Actual closure claims* concern whether our language has some particular closure property that should be reflected in the theory of truth. *Possible closure claims* concern whether languages can be defined that have certain closure properties. Actual closure claims fall under the descriptive project, while positive possible closure claims fall under the normative project. Possible and actual closure claims have different motivations and a theory or language that verifies the former need not verify the latter. Even if some actual closure claims are false, there could be formally or philosophically interesting theories motivated by analogous possible closure claims.

Some philosophers, such as McGee and Beall, think that strong actual closure claims are true and want a theory of truth compatible with those claims. Other philosophers, such as Field, are concerned only with possible closure claims and think that we should use a theory that demonstrates the truth of those claims, because they think that the closure properties

[49] Philosophers primarily engaging in the normative project do motivate and defend the normative features on which they focus. My claim is that they do not satisfactorily defend their selection of descriptive features.

are important and that the actual closure analogs of those claims are false.

As should be clear from the distinction between normative and descriptive projects, not all theories of truth are employed to the same end. The normative and descriptive projects each bring with them differing yardsticks by which to measure theories of truth. These projects have different commitments and burdens to be discharged. Philosophers often fail to say which projects they are undertaking or to acknowledge the existence of the other projects. Getting clear on the arguments in favor of various constraints on theories of truth connected to Semantic Closure will require paying attention to the distinction between the two projects.

1.6 LANGUAGES AND SEMANTICS

Before beginning my analysis of the broad problem of Semantic Closure, I will distinguish different kinds of languages and indicate how they figure in the discussion of Semantic Closure. I will then present a distinction between two approaches to semantics and show how these approaches are relevant to the broad problem of Semantic Closure.

Theories of truth, and semantic theories more broadly, are theories about some language and in some language. The language a theory is about is the object language. If a theory is about a language L, then the language of the theory is a metalanguage for L. A common part of the demand of Semantic Closure is that the object language and metalanguage of an adequate theory of truth must be identical.[50]

There are three kinds of languages that we should distinguish: formal, interpreted, and natural. Formal languages are sets of strings of some alphabet generated by some construction rules. Interpreted languages are formal languages that are interpreted via some interpretation function or model. A natural language is a language that is used by some linguistic community, and so has many features that are not captured in formal or interpreted languages. Some of these features, such as phonetic properties and linguistic history, are not relevant to theories of truth. Some of them, however, are relevant, such as the lack of a clear

[50] I will discuss the argument for this demand and in detail in §2.4.

demarcation in the conceptual resources available in a natural language.

In most cases, there is a gap between formal and natural languages. The relation between a natural language construction and an analysis in a formal language requires some explanation, and one cannot, in general, read off features of the former directly from the latter. This will be important for questions of Semantic Closure, because many of the objections to theories are based on features, or alleged features, of natural languages.

The broad problem of Semantic Closure naturally arises when considering the significance of semantic theories, or the theories of truth they contain. There are two broad approaches to semantics on which I will focus: model-theoretic and truth-theoretic. There are other notions of semantics, but I will not discuss them. The two on which I focus cover the views of the philosophers I will discuss.

A *model-theoretic semantics* defines a class of models and interpretation functions from languages to the models. The Tarski structures of classical first-order logic are used in one sort of model-theoretic semantics. I am using the term "model-theoretic semantics" broadly to include such things as Kripke structures with domains and algebraic models.

I will briefly present a toy model-theoretic semantics for a language. It specifies the interpretations of the basic predicate symbols, function symbols, constants, and variables relative to a domain D of objects.[51] The semantics specifies the conditions under which an atomic formula is true relative to an interpretation, and it recursively specifies the conditions under which complex formulas are true. Given an assignment s of values to variables and an interpretation function I for the predicates, functions and constants, we recursively define a function $Val_{I,s}$ mapping all formulas to the set V of truth values.

- If t is a constant or function symbol, $Val_{I,s}(t) = I(t)$.
- If x is a variable, $Val_{I,s}(x) = s(x)$.
- If $f(t_1, \ldots, t_n)$ is a term then

$$Val_{I,s}(f(t_1, \ldots, t_n)) = I(f)(Val_{I,s}(t_1), \ldots, Val_{I,s}(t_n)).$$

- If $t = t'$ is a sentence, then
  - if $Val_{I,s}(t) = Val_{I,s}(t')$, then $Val_{I,s}(t = t') = \mathbf{t}$, and

[51] The following is adapted from Gupta and Belnap (1993, 40-41).

– otherwise $Val_{I,s}(t = t') = \mathbf{f}$.

- If $F(t_1, \ldots, t_n)$ is a formula, then

$$Val_{I,s}(F(t_1, \ldots, t_n)) = I(F)(Val_{I,s}(t_1), \ldots, Val_{I,s}(t_n)).$$

- If $\sim A$ is a formula, then $Val_{I,s}(\sim A) = \sim Val_{I,s}(A)$.
- If $A\&B$ is a formula, then $Val_{I,s}(A\&B) = Val_{I,s}(A)\&Val_{I,s}(B)$.
- If $\forall xA$ is a formula, then $Val_{I,s}(\forall xA) = glb(\{Val_{s[d/x]}(A) : d \in D\})$.

In the last clause, $s[d/x]$ is the assignment such that $s[d/x](y) = s(y)$, if $y \neq x$, otherwise $s[d/x](y) = d$. If there are other primitive logical connectives, then further clauses will be given for evaluating them.

Model-theoretic semantics is the paradigm of a semantic theory. It provides interpretations of all vocabulary and it can draw many distinctions among the statuses of the sentences of a given language. It has other uses as well, such as use in soundness proofs and in consistency proofs.

A *truth-theoretic semantics* is a semantics in a more holistic sense. It attempts to explain or characterize meaning in terms of truth, which is axiomatized by the set of T-sentences for a particular language. The axiomatization may be broadened to include the quantified forms of various semantical laws, but at its core, it is the set of T-sentences. Davidson's theory of truth is the most famous to take this approach.

The broad problem of Semantic Closure can arise from the relation between these two approaches to semantics. Consider models of the strong Kleene fixed-point theory. They classify sentences into three sets: those assigned **t**, those assigned **f**, and those assigned **n**. This three-way classification cannot be carried out in the object language of the strong Kleene fixed-point theory. All three categories seem to be legitimate semantic categories, so the inability of the fixed-point theory to make this classification in its object language is a deficiency, according to many philosophers.[52] More broadly, truth-theoretic semantics do not seem to capture all of the important semantic information and classifications found in model-theoretic semantics, so there is a gap between the model-theoretic semantics and the truth-theoretic semantics.

[52] See Field (2008, 72-73), for example.

In the literature, one can find two flaws attributed to model-theoretic semantics. First, it seems to deliver only model-relative versions of truth and other semantic concepts.[53] Proponents of some forms of closure are interested in real truth, which is not equivalent to any model-relative notion. The other flaw is that the domains of the models are sets.[54] Models exclude from their domains certain collections of objects. Class models could be used, but these would exclude other, larger collections from their domains. Excluding some collections from the domains creates doubt that the quantified sentences evaluated as true or as valid according to the models are really true or valid.

Truth-theoretic semantics is supposed to be capable of delivering a theory of real truth, because it does not detour through models and model-relative notions. The truth predicate is axiomatized directly. There is, however, a problem with truth-theoretic semantics. It does not make sufficiently many distinctions for the classification of sentences. There is a gap between the ways a model-theoretic semantics can classify sentences of a theory and the ways a truth-theoretic semantics can. A question that concerns proponents of the broad problem of Semantic Closure is whether we can obtain a theory that avoids the problems of model-theoretic semantics while also avoiding the problems of truth-theoretic semantics.

That completes the set up of the preliminaries. The two kinds of semantics will show up repeatedly in the analysis, as I look at claims made concerning the role of truth in semantic theories, especially semantic theories for natural languages. The distinction between the normative and descriptive projects will help focus my discussion and it will be useful in evaluating arguments. With the preliminaries in place, I now turn to the analysis of the broad problem of Semantic Closure, beginning with the hierarchy problem.

1.7 HIERARCHIES

The hierarchy problem is the problem of developing a theory of truth that is free of semantic hierarchies of any kind. Solving this problem has, to many, seemed crucial to the adequacy

[53] See Horsten (2011, 20-22).
[54] See Field (2008, 43-46)

of a theory of truth for natural language. The mere presence of hierarchical concepts in theories of truth has led some philosophers to object to those theories. The Tarskian theory is the paradigm of a hierarchical theory, but it obscures philosophical issues connected to the hierarchies, because the hierarchy of Tarskian truth predicates frequently goes along with other characteristic features, such as restrictions on truthful application of the predicates and the use of additional, distinct metalanguages. Each of these features could be the source of philosophical problems, rather than the hierarchy itself. In this section, I will argue that it is not the mere presence of hierarchies that generates a problem. I will do this by looking at two areas in which hierarchies are found: arithmetic and set theory. I will use the discussion of these areas to help isolate the problem with semantic hierarchies.

1.7.1 Arithmetic

There are hierarchies in the domain of arithmetic. These hierarchies are well-studied and philosophically benign, or at least do not put pressure on the adequacy of arithmetic theories in the way that semantic hierarchies are supposed to put pressure on semantic theories. My task in this section will be to get clearer on this difference.

In formalized arithmetic, we find at least three sorts of hierarchies. There are hierarchies of provability predicates, of consistency statements, and of arithmetical theories. These do not arise within the single system of PA, but they arise once we ask about extending PA in some way to fill the gaps in provability.

The first-order axioms of PA do not decide all sentences in the language of arithmetic. PA cannot classify all of its own sentences as provable or disprovable, even though it can classify some. It cannot prove its own consistency. If one wants to prove the consistency of PA, one can add new axioms, such as a consistency statement, Con(PA). Obviously, the result of adding Con(PA) to PA, PA^+, can prove the consistency of PA, but PA^+ cannot prove its own consistency. A consistency statement for PA^+ can be added to yield PA^{++}. This process can be repeated, and at no point in the progression can the resulting theory prove its own consistency.

The sequence of extensions of PA obtained by adding consistency statements as axioms

results in a hierarchy of arithmetical theories and provability predicates. Each of these provability predicates can be used to classify more sentences of the language of arithmetic as provable or refutable, but none of them classify all the sentences of the language in terms of those predicates.

The unprovability of the consistency of the arithmetic theory "from within" is not generally viewed as a problem. It is viewed as simply a fact of metamathematics. The requirement to go further up a hierarchy of arithmetic theories to prove consistency is not taken to be a philosophical problem.[55] Similarly, the failure of the arithmetical theories to classify all arithmetical sentences as provable or refutable does not create a philosophical problem of the same sort as the failure of a theory of truth to classify its sentences. The hierarchies of arithmetic are philosophically benign, while semantic hierarchies supposedly are not. There is a supposed philosophical difference between the arithmetical case and the semantic one.

In the arithmetical case, we study a fixed, formal language. Changing the logic from one extension to the next is not an option, and the semantics is fixed between extensions. There is a precise specification of what is being added in each extension, and theorems can be proved relating the extensions. In the semantic case, the object of study and analysis is the set of colloquial semantic notions. Self-referential reasoning that leads to theorems in the arithmetic case, often leads, in the semantic case, to paradoxes, the appropriate response to which is not clear. Changing the logic is an option, as is altering the semantics of the language.

We can see from this discussion of arithmetic that the mere presence of hierarchies is not a problem. We can continue the comparisons of hierarchies by looking at the set-theoretic case.

[55] I should mention that Priest provides a dissenting voice. He has argued that the incompleteness theorems provide reason to accept an inconsistent, non-trivial, complete arithmetic. See, for example, Priest (2006, 231-236).

There are philosophical issues generated by the incompleteness theorems, such as questions about new axioms for set theory and challenges for certain philosophical views in the philosophy of mathematics. These seem to be somewhat removed from the current point, so I set them aside.

1.7.2 Sets

ZFC set theory reproduces the hierarchies of arithmetic and generates hierarchies of its own. In addition to the hierarchies from arithmetic, there are ontological hierarchies associated with set theories. Some theories prove more ordinals exist than other theories do. The hierarchy of well-founded sets goes on and on, but there is no set containing the whole hierarchy.[56] There is no way to classify all sets as having some property F or not, without either restrictions on the classification or appealing to resources not supplied by ZFC.[57] For example, one can form the set of all sets in V_ω that have themselves as members and the set of those that do not. One can do the same thing for the whole hierarchy, if one postulates classes, but the problem with classification will reproduce itself at the level of classes. In addition to a hierarchy of set theories, there is an ontological hierarchy of sets.

Not all hierarchies in set theories are problematic. There are philosophically unproblematic hierarchies, such as the Borel hierarchy. These hierarchies will be present in any semantic theory that builds on ZFC. There is no basis to objecting to a set theory on the basis of these hierarchies.

Set theory is frequently viewed as conceptually close to the theory of truth, but there are comparatively fewer objections to the ontological hierarchies of set theory than to the semantic hierarchies.[58] I think that a partial, but non-exhaustive, reason is that ZFC was developed for a normative project, namely providing a consistent or paradox-free foundation for mathematics. The theories of truth with which I am concerned focus on the descriptive project. Philosophers that oppose ZFC on the basis of ontological hierarchies would need to make the case that those hierarchies cast some doubt on the claim that ZFC provides

[56] I will set aside alternative set theories, such as Quine's NF and non-classical theories for this discussion, since ZFC seems to be the set theory most discussed in the literature on truth.

[57] My discussion assumes some plausibility for the so-called naive comprehension schema, but I will not say more on that topic beyond its immediate relevance to issues in the theory of truth. In the future I hope to elaborate the discussion of set theory and its relation to the philosophy of truth developed here. Much has been written on naive comprehension recently. See, for example Restall (1992), Brady (2006), and Weber (2010).

[58] There have been objections to ZFC on the basis of ontological hierarchies. Priest, for example, argues that it is a grave flaw of ZFC that there is no universal set, because there are natural set-theoretic notions that are inexpressible by set theories that generate ontological hierarchies. See, for example, Priest (2006) chapter 2 §3-4. An example of an inexpressible concept is unrelativized set complement.

a paradox-free foundation for mathematics.[59] The problems with hierarchies in the theory of truth stem from other sources, such as difficulties hierarchies create for making sense of certain reasoning patterns involving truth.

A theory of truth T that builds on the language of set theory will contain benign hierarchies. Hierarchies of semantic predicates in T that generate philosophical problems cannot do so solely because they form a hierarchy. The problems must lie in some particularly semantic feature of the hierarchy. I will now turn to a discussion of hierarchies of semantic predicates.

1.7.3 Semantics

Set theories, in particular ZFC, have an intended application that provides them with relatively clear criteria of success. Satisfying these criteria offsets the puzzles that arise from the ontological hierarchies. Semantic theories lack the intended application, as well as the constraint to a single formal language. Consequently, the puzzles stemming from hierarchies appear more pressing in the semantic case.

There is one other feature of semantic hierarchies that philosophically distinguishes them from the ontological hierarchies of set theory. Many semantic notions, such as truth, have pre-theoretic counterparts that those in set theory lack. The concept of a set has some non-technical motivation, but ZFC is used primarily for the normative project of providing a foundation for mathematics. As such, there is less impetus to preserve the intuitions surrounding the intuitive notion. It is hard to see how to preserve these intuitions if the semantic notions are split into hierarchies. There are, of course, some semantic concepts that do not have strong pre-theoretic counterparts. Truth and validity do, while more technical notions, such as determinacy and designatedness, do not. There is less reason to require that the latter not be split into hierarchies.

There are philosophers that think that all infinite hierarchies of semantic notions are philosophically on a par, such as Koons in the following quotation.

> Once the revision theory is applied to the basic notion of categoricalness, a new concept

[59] I am glossing over some important topics in the philosophy of set theory. I think that the point that I am making is relatively uncontroversial.

> of hypercategoricalness is needed in a new metalanguage in order to provide an adequate account of the semantics of first-order categoricalness. This new concept yields yet new paradoxes, and so on, ad infinitum, the Tarskian hierarchy lives![60]

As I understand the argument of the quotation, Koons relies on the unarticulated premiss that all hierarchies or sequences of semantic notions are philosophically equivalent. I will argue that Koons' assumption is false.

Tarskian truth predicates form a hierarchy in a clear way. On a standard way of defining the Tarskian hierarchy, truth predicates lower in the hierarchy do not have in their extensions sentences that use truth predicates higher in the hierarchy.[61] The particular example supplied by Koons provides an important contrast. The categoricalness predicate can apply truthfully to sentences in which it is used. Truth and all categoricalness predicates can be applied truthfully to sentences that use predicates from any level of the hierarchy. Further, the predicates do not build on earlier ones in a straightforward way, as some truths are not categorical.[62] There is a sense in which categoricalness is on a different level of some sort than truth, so there is a kind of a semantic hierarchy. The revision operator for categoricalness is defined over revision sequences for truth, but the revision operator for truth does not look at complete revision sequences for categoricalness. I conclude that not all alleged semantic hierarchies have the same structure.

The important point is that the standard way of defining the Tarskian hierarchy imposes a *rigid hierarchy* while the revision theory imposes a *relaxed hierarchy*. Relaxed hierarchies allow the predicates from all over the hierarchy to apply truthfully to sentences that use predicates from further up the hierarchy. In this sense, the hierarchy of the revision theory is not merely the Tarskian hierarchy under a different name. Theories of truth with relaxed hierarchies can accommodate the self-referential reasoning of Nixon-Dean-style cases, while those with only rigid hierarchies cannot.

Some philosophers, such as Priest, would object, at this point, that my distinction between rigid and relaxed hierarchies misses the point about what is philosophically objectionable about semantic hierarchies: they produce *expressive incompleteness*. The expressive

[60] Koons (1994, 621)

[61] Some versions of the Tarskian approach, such as that of Burge (1979) do not face this problem.

[62] There are some subtle issues involved in this claim. See Gupta and Belnap (1993, 234) for an explanation of this.

incompleteness comes from the lack of a predicate for saying that a sentence is true at some level of the hierarchy, in the Tarskian case, or that a sentence is categorical at some level, in the revision theory case. The predicates of the semantic hierarchy are not closed under combinations of arbitrary subsets of the hierarchy.

The philosophers that oppose hierarchies view each hierarchy as situated within a single formal language.[63] On this view, a philosopher that uses a theory of truth that contains a semantic hierarchy is in some way committed to being able to make sense of talk about the hierarchy. Philosophers may want to talk about the whole hierarchy, and Priest thinks that such talk must not be prohibited.

There are three issues here that I need to pull apart. The first issue is why expressive incompleteness is bad. The second is what problems, if any, arise from having a semantic predicate in a distinct metalanguage. The third is whether and in what sense the semantic predicates of the hierarchy must be closed under arbitrary combinations. I address only the first issue here. I will postpone discussing the second and third issues until §2.4 and §2.2, respectively.

Priest thinks that the expressive incompleteness is a problem because he is engaged in a descriptive project. He thinks that natural languages do not exhibit any sort of expressive incompleteness with respect to semantic concepts, particularly not with respect to hierarchies since the phrase "at some level of the hierarchy" expresses the problematic concept. Therefore, if one's theory of truth does exhibit an expressive incompleteness, then that theory is not adequately capturing the target semantic notion.

Priest's claim that natural languages exhibit no expressive incompleteness is not plausible when we distinguish stages of a language from languages "at the end of development." Understanding the former is a goal of the descriptive project while understanding the latter is not. Language stages may exhibit expressive incompleteness, but this expressive incompleteness will not be a problem. Theories of truth may exhibit expressive incompleteness, but whether it is problematic will need to be argued on a case by case basis, depending on which concept is inexpressible.

There is another objection based on expressive incompleteness that is worth considering.

[63] I take Priest as the primary exemplar of such a philosopher.

Fitch voiced it explicitly about the ramified theory of types.[64] The objection says that if one adopts the restrictions advanced in the theory of types, such as the ban on some kinds of impredicativity, then the theory itself will become unstatable. The theory of types is a theory about all types, but there is no type of all types. A similar criticism can be made of the Tarskian theory of truth.

To respond to this objection, we need to ask what claims are being made about the theory in question. Is the theory of types put forward as a theory of all concepts or as something more restricted? If the former, then the objection holds. Otherwise, the objection is saddling the theory with commitments it did not adopt. Similarly, is the Tarskian theory of truth put forward as a general theory of truth or semantic theory for all languages? If not, as when the theory is put forward as a way to generate a theory of truth for a given language, then it is not clear that the objection has any bite. The objection does not work against arbitrary theories of truth with hierarchies. It works against those with particular philosophical commitments not shared by all theories.

Semantic hierarchies are importantly different from non-semantic ones, such as those found in arithmetic. We can see that not all semantic hierarchies are on a par, as some are rigid and others not. Priest's main objections to semantic hierarchies can be split into three parts, and the immediately relevant one is not convincing. We will turn now to our discussion of semantic self-sufficiency.

1.8 SELF-SUFFICIENCY

In this section, I will present Gupta and Belnap's criticisms of semantic self-sufficiency.[65] McGee is the biggest proponent of semantic self-sufficiency, which connects a semantic theory to understanding its object language.[66] Before presenting the criticisms of semantic self-sufficiency, I will discuss the canonical formulation of semantic self-sufficiency found in Gupta and Belnap (1993). I will then present Gupta's response to an argument against the revision

[64] See Fitch (1946).

[65] I will sometimes use the phrase "self-sufficiency," dropping "semantic." I will use them interchangeably.

[66] This description is particularly accurate in the case of McGee (1991).

theory of truth that uses another kind of semantic self-sufficiency.

Gupta and Belnap gloss the problem of semantic self-sufficiency as "the problem of explaining how a language can be understood from within."[67] This gloss ties the problem to a cognitive property, understanding a language. On this construal, semantic self-sufficiency is a feature of semantics that could shed light on linguistic competency, or at least competency with the concepts involved in the theory of truth. Gupta and Belnap think that appeals to universality, a kind of unboundedness property, motivate semantic self-sufficiency.[68] They say the following.

> Hence, it may be argued, whatever resources may be necessary for understanding a natural language, they can all be expressed within it. The fundamental problem posed by the paradoxes, on this view, is to show how this universality is possible. And theories of truth and paradox that rely on richer metalanguages contribute little to its solution.[69]

If this is correct, then the proponents of semantic self-sufficiency have failed by not explaining the existence this sort of universality. As we will see, however, universality is not the main issue involved in self-sufficiency.

An immediate problem with Gupta and Belnap's gloss is that it relies on an unexplained metaphor of understanding a language from within. Gupta and Belnap do not mean that agents understand a language and were instructed solely in that language. Neither do they mean that the agents understand only a single language.

As an interpretation, I offer the following definition of semantic self-sufficiency, which has two parts. The first part is the *formulation constraint*: a semantic theory about a language $\mathscr{L}$ could be formulated in $\mathscr{L}$. The second part, the *competency conveyance principle*, says that a semantic theory for a language $\mathscr{L}$ potentially imparts competence with, or understanding of, $\mathscr{L}$. The competency conveyance principle has the formulation constraint as a necessary condition. We stipulate this necessary condition because the objections in which semantic self-sufficiency is used belong to a special class: They deal with understanding the language of the theory of truth by appeal to the theory itself.

The background context of the discussion of semantic self-sufficiency assumes that se-

[67] Gupta and Belnap (1993, 257). The first appearance of the label 'semantic self-sufficiency' is in Gupta and Belnap (1993).

[68] I will discuss universality more in §2.1.

[69] Gupta and Belnap (1993, 257)

mantic self-sufficiency is a property of natural languages, and so any theory of truth for natural language that is engaged in the descriptive project must be compatible with this feature of natural language. If a theory of truth fails to satisfy the formulation constraint, it will not satisfy the competency conveyance principle either. Gupta and Belnap do not question the claim that if natural languages are semantically self-sufficient, then descriptive theories of truth must respect this property. Rather, they argue against the claim that natural languages are semantically self-sufficient. Their arguments are split into two kinds, those directed towards the formulation constraint and those directed towards the competency conveyance principle. I will start with the latter.

The initial criticism of the competency conveyance principle that I will present distinguishes two senses of understanding or comprehensibility that are relevant to evaluations of semantic self-sufficiency and arguments in which it features.[70] These arguments are plausible only when the two senses are conflated. Something often cited in favor of thinking that English is self-sufficient is the comprehensibility of English by English speakers, but there are two possible senses of comprehensibility available. One is the ability to use and grasp the language. The other is the ability to give a systematic semantic theory for the language. In the former sense, the claim is trivial. In the latter sense, the claim is that English speakers can give a systematic semantic theory for their language, but there is no evidence for this claim. For example, one can ride a bike without any problems, although it may be quite hard to state systematically and explicitly what one has to do to ride a bike. Being able to do something does not generally entail that one can formulate principles systematically explaining the activity. The comprehensibility of English by English speakers does not lend any support to the competency conveyance principle, but comprehensibility was the evidence for the competency conveyance principle. Without that support, there is not, as far as I can tell, reason to think that semantic theories do, let alone must, satisfy the competency conveyance principle.

I will now turn to three arguments against the formulation constraint.

[70] This criticism is part of the response in Gupta (1997) to arguments, found in McGee (1997) and Martin (1997). These arguments attempt to provide reason to reject the revision theory of truth on the grounds of semantic self-sufficiency. I will present these arguments in §1.8.3.

1.8.1 First argument

Gupta and Belnap think that semantic self-sufficiency is motivated by a false premiss, namely a form of universality.[71] They do not, however, find any form of it that both supports the self-sufficiency problem and appears to be true of natural languages. They distinguish two forms of semantic universality, the former of which they accept.

∀∃ **GB Universality** For any semantic concept C, there is a language $\mathscr{L}$ that contains its own C-concept.

The corresponding ∃∀ claim is rejected.

∃∀ **GB Universality** There is a language $\mathscr{L}$, such that for any semantic concept C, $\mathscr{L}$ contains its own C-concept.

In addition to thinking that ∃∀ GB universality is not a property possessed by natural languages, Gupta and Belnap say of this version, "we do not even have an abstract conception of what the requisite [language] would be like."[72]

Gupta and Belnap's endorsed form of universality, ∀∃ GB universality, does not support either part of semantic self-sufficiency. ∃∀ GB universality supports the formulation constraint. It may support the competency conveyance principle as well, if there is a tight connection between theories of semantic concepts and language understanding. We would need evidence that natural languages satisfy ∃∀ GB universality in order to use that property to argue for semantic self-sufficiency. Without such evidence, ∃∀ GB universality provides no support for the competency conveyance principle.

In §2.1, I distinguish two different kinds of universality found in Tarski's work. Neither of these, however, supports the competency conveyance principle, as they make no connection to understanding, and neither supports the formulation constraint. I conclude that appeal to universality does not support semantic self-sufficiency, so I now turn to the second argument.

[71] Universality will be the topic of §2.1.
[72] Gupta and Belnap (1993, 258)

1.8.2 Second argument

Gupta and Belnap argue that there is no evidence for semantic self-sufficiency as a property of natural language.

1. Evidence for semantic self-sufficiency is either *a priori* or *a posteriori*.
2. There is no *a priori* evidence for the claim that natural languages must be self-sufficient.
3. There is no *a posteriori* evidence that natural languages must be self-sufficient.
4. Therefore, there is no evidence for semantic self-sufficiency.

The reason for accepting the third premiss is that it is unknown what resources will be required for a full semantic theory for a natural language. Semantics has not developed fully enough to pronounce a verdict one way or the other. This argument does not rule out support being developed in the future.

This argument provides support for an agnosticism about the semantic self-sufficiency of natural languages. It does not need to lead to skepticism that semantic theories for natural languages cannot be developed.

1.8.3 Third argument

Martin (1997) and McGee (1997) press an argument against the revision theory that an understanding of the semantics of natural language requires the formulation of a description of the semantic theory in the target language. Their primary objection is as follows.[73]

1. An adequate semantic description of English is possible.
2. This description must be formulable in English itself.
3. Revision-theoretic semantics for a language can only be constructed in a richer metalanguage.
4. Therefore, revision-theoretic semantics is not suitable for English.
5. Therefore, revision-theoretic semantics does not capture the notion of truth in English.

[73] My presentation closely follows that of Gupta (1997), which seems correct.

Gupta objects to the second premiss of the argument, which is a variation of the formulation constraint, that a language $\mathscr{L}$ modeling a natural language should be capable of expressing its own semantic theory.

To set up the argument, English needs to be idealized as frozen at a particular stage in its development to prevent extraneous concerns arising about the adequacy of a semantic description at different times. Part of the reason for thinking that English can provide an adequate semantic description of itself is that it has a great deal flexibility to create expressions for denoting a range of objects and properties.

Gupta poses a dilemma. Either the expressibility of an adequate semantic description of English is due to this flexibility or it is not. For the first horn, suppose that it is. Then there is no support for the claim that English can express an adequate semantic description of itself. We have assumed fixed conceptual resources, so the flexibility of English cannot support that claim after all. For the second horn, suppose that the semantic self-sufficiency of English is not due to the flexibility of English. In that case, we can suppose that English has fixed conceptual resources. Given that, there is no empirical evidence that English does express an adequate semantic description of itself, and there is no *a priori* support for it either. In neither case is there any evidence for the second premiss.

1.8.4 Conclusions on semantic self-sufficiency

Semantic self-sufficiency should be rejected as a desideratum on theories of truth. Gupta and Belnap's criticism of it are sound. Apart from Martin and McGee, no one else appears to endorse the competency conveyance principle as a desideratum, and, in recent work, even McGee has moved away from it. The formulation constraint is adopted by other philosophers. The intuition underpinning it was articulated by Martin when he said that we have only our language in which to talk about truth.[74] This intuition stems from the closure problem and the metalanguage problem, which issues I discuss in §2.2 and §2.4, respectively.

[74] Martin (1997, 417-418)

1.9 CHAPTER SUMMARY

In this chapter, I have presented the historical background for the broad problem of Semantic Closure. To get clear on the conceptual issues, I distinguished the descriptive and normative projects for theories of truth as well as distinguishing six aspects or subproblems of the broad problem of Semantic Closure. The conclusions of the analysis so far have been primarily negative. Two of the subproblems are not the central issues raised by the broad problem.

In the next chapter, I will examine the remaining problems, with a particular focus on the problems comprising my proposed analysis of the broad problem of Semantic Closure: the classification problem and the metalanguage problem. In the course of the analysis, I will argue for substantive criteria of adequacy for theories of truth.

2.0 VERSIONS OF SEMANTIC CLOSURE

In the previous chapter, I provided historical background on the broad problem of Semantic Closure and distinguished several problems that are involved. I argued against both the claim that the presence of semantic hierarchies is a problem for theories of truth and the claim that semantic self-sufficiency is a good constraint on theories of truth. In this chapter, I will analyze the remaining problems. I will begin by looking at the universality problem in §2.1. I will then turn to the narrow closure problem in §2.2. In the successive two sections, I will focus on the problems I labelled as central to the broad problem of Semantic Closure: the classification problem (§2.3) and the metalanguage problem (§2.4). Finally, in §2.5 I will argue that semantical laws are a necessary condition on any theory of truth to be a legitimate semantic theory.

2.1 UNIVERSALITY

Natural languages are supposed to have a property, universality, that is used to motivate approaches to the paradoxes and some forms of Semantic Closure. While many philosophers mention universality in discussions of Semantic Closure, the concept rarely comes under much scrutiny.[1] In this section I will distinguish two senses of universality in the work of Tarski.[2] Of these, only one can plausibly be used to argue for forms of Semantic Closure. In §2.1.3, I will argue for an open-ended requirement on theories of truth based on considerations of universality.

[1] Simmons (1993) is an exception.

[2] In §2.2, I will discuss a condition deriving from Tarski's work. It is not really a kind of universality, so I will delay discussion of it.

Tarski (1983) cites the universality of natural language as the source of antimonies in natural language, saying: "it could be claimed that 'if we can speak meaningfully about anything at all, we can also speak about it in colloquial language'."[3] Tarski presents this notion and then proceeds to extract some ideas from it and, using those, draws conclusions about theories of truth. There is a germ of truth to his idea, but there is a gap between that germ and its relevance to theories of truth.

Universality is supposed to be an uncontroversial property of natural languages. Tarski does not argue for it. He simply claims it as a commonsense feature of natural language. Tarski famously despaired of giving a theory of truth for natural language, citing universality as a serious obstacle to defining truth for natural languages. The majority of what he says about universality appears while discussing the nature of natural language.

> A characteristic feature of colloquial language (in contrast to various scientific languages) is its universality. It would not be in harmony with the spirit of this language if in some other language a word occurred which could not be translated into it; it could be claimed that 'if we can speak meaningfully about anything at all, we can also speak about it in colloquial language'. If we are to maintain this universality of everyday language in connexion with semantical investigations, we must, to be consistent, admit into the language, in addition to its sentences and other expressions, also the names of these sentences and expressions, and sentences containing these names, as well as such semantic expressions as 'true sentence', 'name', 'denote', etc. But it is presumably just this universality of everyday language which is the primary source of all semantical antinomies, like the antinomies of the liar or of heterological words.[4]

There are two primary ideas concerning universality that I find in the work of Tarski. The first is the idea that all concepts are expressible by natural languages. The second is the idea that natural languages can always be extended with new vocabulary. I will discuss each in turn to figure out what constraints they put on theories of truth.

2.1.1 Expressibility

The first idea concerns expressibility and an initial formulation of the idea is as follows.

Naive Expressibility Natural languages express all concepts that are expressible.

[3] Tarski (1983, 164)
[4] Tarski (1983, 164)

Naive Expressibility brings with it some problems concerning whether all concepts are expressible. We can avoid some these problems by noting that only a particular subset of concepts matter for dealing with a theory of truth, namely the semantic concepts, refining Naive Expressibility to the following.

Expressibility Every natural language expresses all semantic concepts that are expressible.

This is doubtful. Some languages may use different semantic concepts that are not interdefinable with truth. One may use truth and falsity, while another contains semantic expressions not definable in terms of expressions in the first language.

Even if a natural language is not able to express all semantic concepts, that language can, presumably, express all of its own semantic concepts. This is a standard, if frequently undefended, claim that comes up when discussing theories of truth.[5] Other languages may use exotic semantic concepts, which our language need not express. This leads to the following modification of Expressibility.

Restricted Expressibility Every natural language L expresses all of the semantic concepts necessary for the semantics of L.

Restricted Expressibility makes a descriptive claim. It says that every natural language expresses the semantic concepts necessary for that language's semantics, but it leaves open the question of whether natural languages express semantic concepts not necessary for their semantics.

There are two problems with Restricted Expressibility. First, it uses an unanalyzed notion of expression, which is supposed to be a general notion, applicable to natural languages as well as interpreted formal ones. The lack of a developed theory of expression hampers its use in the area of theories of truth. It can be interpreted in a narrow way, for example as definability, but this is not compatible with the broader ways in which it has been used. Expressibility is usually taken to outstrip definability. There are precise notions that we can use with respect to formal theories of truth, but they do not readily translate over to natural languages.

[5] Here is a representative quotation. "A natural language (or a formal language that models that aspect of its behaviour) *can give its own semantics*." (Priest, 2006, 70) The emphasis is mine.

The second problem with Restricted Expressibility is that it tells us little about the concepts that languages must express. It will provide some constraints on theories of truth if some concepts can be established as necessary for an account of the semantics of a particular language.

These two problems can be solved by a detailed theory of expressibility, but none of the philosophers I discuss have offered one.[6] I am not going to develop a theory of expressibility, but I think that aspects of Restricted Expressibility do not need such a theory. Some of the appeal to Restricted Expressibility comes from the idea that there seem to be certain semantic ideas that we either can express or can introduce means of expressing. These cases can, I think, be understood in terms of expanding a language. For these reasons, I will set aside expressibility and move on to discuss the next idea, extensibility.

2.1.2 Extensibility

The second idea concerns extending a language. This idea can be found in the following quotation, which comes just before Tarski introduces the term "universality."

> [The specification of true sentences in terms of structural features] seems to be almost hopeless, at least as far as natural language is concerned. For this language is not something finished, closed, or bounded by clear limits. It is not laid down what words can be added to this language and thus in a certain sense already belong to it potentially.[7]

Tarski points out that natural languages are not "closed" or "bounded by clear limits" and that vocabulary can be added to them without any apparent restrictions. These points point towards an important feature of natural languages: they can be extended. A theory of truth that is supposed to capture natural language will need to be compatible with this idea. There are, however, two ways of understanding this principle, corresponding to two ways of interpreting Tarski's claims. The first way of understanding this expansion is that for every stage of a language and word not in the language at that stage, there is an extension of the language containing that word. This does not require any final stage or language stages that

[6] Work by James Shaw starts to address the second problem. Shaw (2013) discusses some expressibility requirements on theories of truth.

[7] Tarski (1983, 164)

are maximal under the extension relation.[8] This suggests the following principle.

Extensibility ($\forall\exists$) For all language stages E, if there is a predicate P not in E, then there is a language stage E' that contains P and extends E.

With some quantifier manipulation, the preceding principle becomes an $\forall\exists$ claim.[9]

The second interpretation of the Tarski quotation is suggested in the final sentence of the quotation, where Tarski talks about a word belonging to a language potentially. This interpretation reads Tarski as saying that there are language stages that contain all possible vocabulary. Such stages are maximal under the extension relation. A word potentially belongs to a particular stage if it belongs to a maximal extension of that stage. This gives rise to $\exists\forall$ version of Extensibility ($\forall\exists$).

Extensibility ($\exists\forall$) There is a language stage E'' such that for all possible predicates P and all language stages E and E', if P is in E' and P is not in E, then E'' extends E and P is in E''.

This claim is less plausible than Extensibility ($\forall\exists$), although the latter already requires some amount of idealization.[10] There is not a general reason to accept the claim that the space of language stages, or the part under consideration, is arranged in such a way that the set of stages without some predicate P has a maximal element with respect to the extension relation. There is additionally a question of what the range of possible words is.

We can reject Extensibility ($\exists\forall$) while still holding that languages are extensible and make sense of Tarski's comment about words potentially belonging to languages. We can do this by accepting Extensibility ($\forall\exists$) and saying that a word potentially belongs to a language if there is a stage of the language containing that word.[11]

Extensibility is a claim about the existence of languages. On its own, this does not tell us what a theory of truth should be like. This is because of the gap between natural languages and interpreted formal languages mentioned in §1.6. In the next section, I will argue for a

[8] Although this is a loose formulation, the idea of an ordering of one language extending another language should be clear enough for maximality to make sense.

[9] There are analogous versions of Extensibility $\forall\exists$ for operators and terms as well. For simplicity, I will focus on the predicate version and use "predicate" and "word" indifferently in this context.

[10] It requires some idealization to view any of the languages as plausibly having infinitely many extensions.

[11] There are likely some issues that may arise from combining the stages with the right sense of possibility.

way in which Extensibility ($\forall\exists$) is relevant to theories of truth.

2.1.3 Logic neutrality

The extensibility of natural language is relevant for theories of truth engaged in the descriptive project. A theory of truth needs to accommodate the possibility of new vocabulary.

Extensibility ($\forall\exists$) is supposed to capture the fact that our language expands. Our language already has a truth predicate and resources enough for many paradoxes to arise. Given the current state of our language, we are confronted with problems about paradox. We could take a time slice of our language and develop a theory of truth for that time slice. We could take the semantic-free part of that time slice as the base language L and add the truth predicate and the additional semantic vocabulary. From Extensibility ($\forall\exists$), we know that there are extensions of this that contain other logical operators and exotic semantic and expressive vocabulary. These seem to be equally legitimate base languages to which to add truth. An account of truth that did not carry over to at least some of the possible extensions would be inadequate.

We clearly encounter problems when developing theories of truth for very rich languages. However, even had we not begun with such a rich language, we would, plausibly, run into the same or similar problems. I think that looking at a common way of setting up the problem can bring out the issue with Extensibility ($\forall\exists$).

Most approaches to the theory of truth begin with a classical ground language, which is then expanded with a truth predicate.[12] Beall (2009) presents it as a kind of origin myth. Suppose that a community used a natural language in a way that is best regimented with a classical first-order language.[13] Someone, call him "Jones," comes to the community and introduces the truth predicate and syntactic vocabulary to this community. Jones teaches the community to use truth. The scheme of the language, the specification of the semantic values and the way to evaluate the logical vocabulary, after the introduction of the truth predicate is frequently different than before.[14] For example, with the minimal strong Kleene fixed-

[12] See Kripke (1975), Field (2008), Gupta and Belnap (1993), for example.
[13] This can be thought of along the lines of Lewis (1975).
[14] See Gupta and Belnap (1993, 40-44) for details on some particular schemes.

point theory of truth, the language before the introduction is 2-valued whereas afterwards it is 3-valued. Not all approaches to truth require altering the scheme of the base language, but such alteration is illustrated well by the strong Kleene example.

Our imagined community speaks a language that is best regimented as classical before they learn how to use the truth predicate. There are extensions of this language that contain other logical connectives. A few precocious users of this language could introduce new logical operators and change the basic scheme of the language. Perhaps they distinguish a kind of negation that validates excluded middle in a 3-valued context and one that does not, or perhaps they want a conditional that remedies deficiencies they find in their classical material conditional. Their linguistic behavior will be best modeled by one of the extended languages. For concreteness, suppose our community adopts a language based on the strong Kleene scheme with choice and exclusion negations.

Returning to Beall's myth, Jones comes to visit this community and introduces the truth predicate to them. Jones cannot get them to adopt a strong Kleene fixed-point theory of truth, because the inductive construction will not reach a fixed-point in a 3-valued setting due to the non-monotonicity of exclusion negation.

The community's starting point does not mean that they are barred from adopting a fixed-point theory of truth. They can adopt one, but it will not be the strong Kleene fixed-point theory.

There are languages with different expressive resources and schemes. Extensibility ($\forall\exists$) says that there are extensions of the community's language that contain some of these expressive and logical devices. These extended languages are, it seems, legitimate starting points for adding the truth predicate, so Extensibility ($\forall\exists$) provides some reason to consider a range of base languages.

Having seen how Extensibility ($\forall\exists$) works in a simple case, we can return to the case of rich languages similar to the ones that we use. We have a particular time slice of the language for which we can develop a theory of truth. There are logical and semantic resources that may be added to the languages. These are also legitimate starting points for developing a theory of truth. Ignoring or barring any of these languages while developing a theory of truth seems to be an *ad hoc* response. This is because those languages are ones that we

could use, as are the earlier ones.

My arguments show that an adequate account of truth for natural language needs to be *logic neutral* in the following sense.

Logic Neutrality An account of truth is *logic neutral* iff the account can be applied uniformly to a range of base languages, each of which contains a range of logical and semantic resources, and the account produces an adequate theory of truth regardless of which base language from the range is chosen.

Let us look at some examples. A theory that works well only when no negation is in the language is clearly inadequate. The strong Kleene minimal fixed point approach to truth is not logic neutral either, as it does not work if the base language contains non-monotonic logical operators.

As it stands, Logic Neutrality has many details to be filled in. First among these is what the ranges of base languages and of logical resources are supposed to be. There are, after all, constructions in English that are well rendered formally using second-order logic, general quantifiers, sortal quantifiers, plural quantifiers, and term-forming operators. There are base languages that contain these resources, so, if this argument has been sound, adequate theories of truth will need to work with those languages.

We might initially start by saying that the range of languages and logical resources is arbitrary. I want to begin with a weaker base line while maintaining the motivating spirit. To begin, we restrict the range of the base languages to languages that have a well-founded syntax and, at least initially, a first-order structure. There do not seem to be any restrictions needed on the number of values in the schemes for the starting base languages. Schemes with infinitely-many values may have truth functions with infinitely-many arguments that are not captured by any finite truth functions. I am not sure whether neutrality requires that an account of truth work with such infinitary operations, but it seems plausible that an account of truth should be compatible with finitary truth functions. These restrictions can be motivated by Extensibility ($\forall\exists$), so they are ways in which the theories of truth respect the extensibility aspect of natural language.

Logic Neutrality is a requirement on theories of truth stemming from Extensibility ($\forall\exists$).

Given a base language, which is one that regiments a way we might use our language, Extensibility ($\forall\exists$) says that there are possible extensions of it that contain various logical and semantic resources. Theories of truth should work with these possible extensions as well as the particular starting point. Logic Neutrality is not yet precise, however, as there are still questions about the range of logical and semantic resources that need to be settled. I will not make the notion more precise, as doing so is beyond the scope of this dissertation. Two basic features of the notion will be sufficient for my purposes, namely that it is a consequence of Extensibility ($\forall\exists$) and that it will rule out a range of theories of truth.

2.1.4 Conclusions on universality

I have examined universality as it is found in the work of Tarski. I distinguished two forms of universality found in Tarski's work: Expressibility and Extensibility.

Considerations of Extensibility lead to the suggestion that theories of truth should be compatible with base languages containing a variety of logical resources, which is the constraint of Logic Neutrality. There are further details to work out on the exact extent of Logic Neutrality. Extensibility ($\forall\exists$) seems to be an important feature of natural language, one that descriptive theories of truth should reflect. Respecting Logic Neutrality is required by the descriptive project as well, since Logic Neutrality is a consequence of Extensibility ($\forall\exists$). In order to satisfy Logic Neutrality, a theory of truth must be neutral with respect to the logic of the base language, and not all theories of truth are. This feature of the descriptive project has not been remarked upon much.

I will now turn to the narrow closure problem.

2.2 CLOSURE

One of the guiding intuitions behind the broad problem of Semantic Closure is that the language of the theory of truth should be closed under certain operations or constraints relating to the semantic vocabulary. In this section, I will discuss the narrow problem of

closure, namely with what sort of closure requirements a theory of truth must be compatible. I will present the constraints on closure of the syntactic theory that I find in Tarski's work. I will then turn to Priest's discussion of closure. I will present Priest's formulation of one kind of closure condition, which I will generalize. I will then look at some of Priest's claims relevant to closure under the addition of semantic vocabulary.

2.2.1 Syntactic closure

The first closure condition is a condition on the background syntactic theory. This condition is endorsed in Tarski (1983).

Syntactic closure For any sentence A of the language E, E contains a name of A.[15]

Syntactic closure is a property that natural languages have. Syntactic closure is a property of languages, but it can be extended to theories of truth by saying that a theory has syntactic closure if its language does. It is a reasonable constraint to put on an adequate theory of truth that it have syntactic closure. The truth predicate applies to terms, so there must be some syntactic resources in the language providing names of sentences.

Tarski's formulation can be strengthened. If the theory of truth is to be a part of a rich semantic theory, then the principle should be strengthened to the following.[16]

Satisfaction Readiness A language E is satisfaction ready iff E contains names of all of its sentences, names for predicates of each arity, and a satisfaction predicate.

This strengthens the syntactic theory so that the theory of truth can use satisfaction in addition to the truth predicate. Tarski's use of Satisfaction Readiness requires that semantic predicates, such as denotation, be in the language, and so names for those will be in the language as well.

There is an addition to Satisfaction Readiness that is worth making. Nothing said so far requires that there be ways of manipulating terms for sentences in a way that mirrors

[15]Tarski used the notion of structural descriptive names for sentences. Other standard options include quotation names and numerals, via a coding. Other singular terms will work, including definite descriptions and deictic terms, but my focus will be on languages that contain either quotation names or use a coding scheme.

[16]Tarski (1969, 69) endorses this version. This is also endorsed by Simmons (1993).

logical structure, in the sense that if a language has a sentence $C(A_1, \ldots, A_n)$, where C is a connective C and $A_1, \ldots, A_n$ are sentences, then there is a function symbol for a function f where

$$f(\ulcorner A_1 \urcorner, \ldots, \ulcorner A_n \urcorner) = \ulcorner C(A_1, \ldots, A_n) \urcorner,$$

the right-hand side of which names the sentence $C(A_1, \ldots, A_n)$. Natural languages are able to manipulate names of sentences and their parts and they contain names for predicates and terms. This motivates a requirement that theories of truth give the correct verdicts when combined with languages that can manipulate names of sentences and formulas into names of other sentences. A theory of truth that was only adequate for languages that could not manipulate sentence names would not capture the target notion of truth.

2.2.2 Semantic closure

Priest (1984) describes a closure condition, which he calls *semantic closure*, that he thinks is necessary for the adequacy of a theory of truth as a semantic theory: "a semantically closed theory/language being one which can adequately express its own semantic concepts."[17] Priest takes satisfaction to be the primitive semantic concept and uses that to define truth and denotation. He gives the following definition.

> A theory is semantically closed (with respect to its satisfaction relation) iff (i) for every formula with one free variable ϕ, there is a term a_ϕ, its name, (ii) there is a formula with two free variables $Sat(x, y)$ such that every instance of the scheme
>
> $$Sat(t, a_\phi) \leftrightarrow \phi(v/t)$$
>
> is a theorem, where t is any term, ϕ any formula with one free variable v, and $\phi(v/t)$ is ϕ with all occurrences of 'v' replaced by 't'.[18]

Priest's definition breaks into two parts. The first is the closure part, which requires that for every formula with one free variable, there is a name of that formula. The second part is what I take to be the adequacy part. The opening gloss of semantic closure is that of adequately expressing semantic theory. A theory adequately expresses its satisfaction relation if it entails all the instances of the above biconditionals.

[17] Priest (1984) p. 118
[18] Priest (1984) p. 118

Priest's semantic closure will be inadequate if there are semantic concepts in the language not definable in terms of truth and satisfaction. Modifying the definition to take this into account yields the following.

Generalized Semantic Closure A theory is *semantically closed* with respect to some semantic concepts iff

- for all n, each n-ary predicate ϕ has a name a_ϕ, and
- the theory validates all the sentences of the adequacy conditions on the semantic concepts.

For the particular cases of satisfaction and truth, the adequacy clause could be rephrased to say that the theory validates all instances of the satisfaction biconditionals and the T-sentences.

The adequacy part is the interesting part of the definition. It is, I think, a good requirement on theories for two reasons. First, since Tarski's work, many philosophers have viewed the T-sentences and the satisfaction biconditionals as presenting the core of these two concepts, if not exhausting their content. A theory that invalidates some of these biconditionals needs some defense to indicate how it is a theory of the target notion of truth. Second, there is some philosophical tension created by the theory invalidating principles that it holds to be good, in particular, conditions by which we want to evaluate the adequacy of the semantic vocabulary. Different views may be able to relieve some of this tension by pointing to adequacy conditions stated in a metalanguage and providing a justification for the failure in the object language. The reasonableness of such a justification depends on the overall philosophical view of which it is a part.

Generalized semantic closure places few constraints on the logical resources of the theory or language. One could have a language that lacked negation, and this could be semantically closed. It does not require that much be in the theory at all. The closure condition does not require that there be any function or relation symbols in the language that capture basic facts about the syntax of the language. Generalized semantic closure applies as well to weak or impoverished languages as to rich languages.[19]

[19]There are interesting questions about how truth behaves in impoverished languages, as discussed in Gupta (1982). My focus here is on rich languages, but the behavior of truth in languages that are poor in

Generalized semantic closure places few demands on semantics, but it does cut against some theories of truth. For example, neither the strong Kleene fixed-point theory of truth nor the revision theory satisfy the requirement of generalized semantic closure. It seems to be a reasonable demand on any theory that claims to be semantically closed.

2.2.3 Predicates

In §1.7.3, I raised and deferred a question brought up by one of Priest's arguments. The question was whether and in what sense the semantic predicates of a semantic hierarchy must be closed under arbitrary combinations. I will respond to this question now.

To begin, the question is not really about arbitrary combinations of predicates. It is better put in terms of intuitively expressible concepts. The hierarchy of Tarskian truth predicates is the paradigm example for this objection. The concern is not with arbitrary combinations of truth predicates, but rather with the concept of being true at some level. This seems to be a concept that we have a grip on and it is intuitively expressible by someone with an understanding of the Tarskian theory of truth. It is expressible and not definable, because there is no way, on the standard Tarskian theory, of quantifying over the indices of the Tarskian truth predicates. Being true at some level does seem to be a legitimate concept, so there is some force to Priest's complaint that the Tarskian theory does not permit it.

The problem seems to turn primarily on two things: a worry about evaluating the predicates and a worry about expansion with vocabulary. I will take these in turn.

In §2.1, I argued that there is good reason for requiring a theory of truth to be compatible with expanding the language with new logical and semantic vocabulary. One would expect that a theory of truth should be compatible with expanding the language by a kind of "maximal" semantic predicate, such as "true at some level". If a theory is incompatible with this, it would be a strike against that theory.

The problem seems to have the most force against the rigid semantic hierarchies, hierarchies whose predicates do not apply to sentences located further up the hierarchy. A "true at some level" predicate would, presumably, need to apply to sentences in which it figures,

syntactic and semantic resources is an important question.

so it would quickly violate the basic Tarskian idea of a rigid hierarchy of truth predicates. For relaxed hierarchies, hierarchies in which there is no restriction on application, it is not clear that there is a problem with expanding the language with the new vocabulary. This leads to the worry about evaluation.

The worry about evaluating predicates is that the predicates introduced would not be evaluated as expected. In particular, the maximal semantic predicates would not be maximal. For example, in the revision theory, viciously circular sentences exhibit a kind of semantic pathology. A maximal semantic predicate would classify all such sentences, but it could be used to create new liar-like sentences that would exhibit a new kind of semantic pathology.

A tentative response is that, while a predicate could be introduced, it may generate new semantic concepts or categories. In such a case, there would not be a guarantee that the maximality aspect would be maintained. As an analogy, consider the metaphor of generating the finite ordinals with ordinal addition. If one stipulates that there is an ordinal bigger than all of those, a greatest one, from a standpoint considering only the finite ones, then one will have ω. The principles of ordinal addition lead to larger ordinals, so the maximality stipulation cannot be maintained with ordinal addition.[20]

An alternate response is that certain kinds of maximal semantic concepts cannot be introduced into a language, contrary to initial appearances. It is a failure along the lines of the failure of denotation of the expression "the greatest natural number." Such a response is not available in all cases. For example, it seems implausible in the Tarskian case, because the levels of the Tarskian hierarchy are supposed to be aspects of a single, intuitive concept, namely truth.[21] When the semantic hierarchy does not correspond to a concept that has an apparent informal or intuitive counterpart, this sort of response will, I think, seem more plausible. The categoricalness predicate of the revision theory is an example of a semantic predicate without an apparent intuitive counterpart. Depending on the philosophical view involved, operators akin to Field's determinateness operators may be as well.

Based on my arguments from earlier in this chapter, there is reason to expect a theory

[20] I suspect that a logician who favored paraconsistent set theory would argue that the phenomenon I am discussing points to a fixed-point of ordinal addition, an inconsistent ordinal, so to speak. I am unsure how to respond to this point here.

[21] This depends on a particular view about the purpose of the Tarskian theory that Tarski did not seem to have.

of truth to be compatible with a range of additions of semantic vocabulary built, in some sense, from semantic vocabulary already in the theory. Failures on this front are, sometimes, good objections to theories, but they are not good objections in all instances.[22] There are some views of truth and related semantic concepts that can plausibly respond to Priest's arguments, but these responses depend on the details of the view in question.

2.2.4 Conclusions on closure

In this section I presented two kinds of closure conditions on theories of truth, syntactic and semantic. The syntactic conditions are well motivated by appeal to the descriptive project. Requiring that the language for a theory of truth contain names for formulas is reasonable.

The condition of generalized semantic closure, likewise, receives some support from the descriptive project. The clause stating that the theory of truth must validate the adequacy conditions for the semantic concepts it contains does rule out some theories. That clause says that theories of truth must validate the T-sentences, and this will rule out theories such as the standard revision theory or the standard strong Kleene fixed-point theory.

I will now turn to the classification problem.

2.3 CLASSIFICATION

The classification of sentences into different categories is a central component of Semantic Closure. In this section, I will look at views on the classification of sentences into semantic categories. I will present the views of Kripke, Field, and Beall. I will formulate a principle specifying the amount of semantic classification that is required by a theory of truth to be sufficiently semantically closed for the descriptive project. I will also argue in §2.3.4 for the importance of a class of semantic predicates, the diagnostic predicates.

[22] See Shapiro (2011) for a good discussion of this point.

2.3.1 Kripke

As I said in §1.3, the broad problem of Semantic Closure can be traced, in part, to Kripke (1975). Kripke said that one had to ascend to a metalanguage in order to say things about some sentences on the strong Kleene fixed-point view.

> [T]here are assertions we can make about the object language which we cannot make in the object language. For example, Liar sentences are *not true* in the object language, in the sense that the inductive process never makes them true; but we are precluded from saying this in the object language by our interpretation of negation and the truth predicate.[23]

The liar sentence will never be put into either the extension or the anti-extension of the truth predicate. To put it in a different notation, the liar receives the semantic value **n** in all fixed-points. There is no way to say this truly in the object language. There is no way to classify the semantic status of the liar in the standard strong Kleene fixed-point theory of truth.

One way of classifying a liar sentence would be as having semantic value **n**. A predicate interpreted in this way, whose extension is the set of sentences in neither the extension nor the anti-extension of the truth predicate cannot be added to a strong Kleene fixed-point theory of truth.[24] Another way is to use the truth predicate to express the concept of a truth value gap as follows.

$$\sim(Tx \vee T(\dot{\sim}x))^{25}$$

This complex predicate, when applied to a liar sentence, will be evaluated as **n**. This may be viewed as a failure of the material adequacy of this complex predicate.

There is a strong feeling that there is something amiss with liar sentences, but the strong Kleene fixed-point theory does not offer any way to diagnose this in the object language. There is a failure on the part of the theory to classify sentences of the theory.

The fixed-point model is defined in a set-theoretic metalanguage for the theory. The model classifies all the sentences of the theory into three different sets. If the models correctly

[23] Kripke (1975, 79-80) Emphasis in the original.

[24] I am going to set aside the option in which another truth value is added for the inductive construction.

[25] The symbol $\dot{\sim}$ stands for a function that maps a name of a sentence to a name of the negation of that sentence. I am assuming that there is only one negation in the language. If there are multiple negations in the language, an additional function will be needed for each negation.

represent the semantics of the truth predicate, then the theory of truth fails to represent the semantics of its truth predicate adequately in its object language.

Kripke thinks that it is a problem that the object language theory does not classify the liar, but it is not clear why he thinks this. We could, following Halbach and Horsten (2006), deny that the liar should be classified, rejecting the notion of truth used in the classification as illegitimate, not real truth. This view takes the theory of the fixed-point to capture real truth and concepts not definable in terms of that notion of truth and the attendant strong Kleene logic as, likewise, illegitimate. One can do this, but doing so gives up on the descriptive project. There is, as I said, a strong intuition that there is something amiss with the liar sentences. This intuition is, I take it, a large motivation for Field's classification of the liar as defective in some way.

Kripke cites one of the features of the fixed-point theory as allowing the reconstruction of intuitive arguments and concepts involving self-referential truth. To reject the claim that the liar is in some way defective is to reject many of the classifications offered by the models of the fixed-point approach. Kripke's adoption of the descriptive project provides some motivation for classifying the liar in some way.

We can divide the classification of sentences into two sorts. The first is the attribution of semantic value. Predicates that are supposed to capture or represent certain semantic values are predicated of sentences. The strong Kleene fixed-point theory fails at representing semantic values, since there is no predicate whose extension is all and only the sentences evaluated as **n**. For the strong Kleene theory, the true sentences and the false are classified with the predicates Tx and $T(\sim\!\!\cdot x)$, respectively.

The second sort of classification of sentences is *diagnostic*. This is the attribution of statuses to sentences that do not directly correspond to truth values of the model theory but that do provide information about the logical behavior of the sentences. For example, the pathologicality predicate is true of all and only sentences that are in no fixed-points.

The distinction between semantic value classification and diagnostic classification is important, and I will discuss it further after presenting the views of Field and Beall.

2.3.2 Field

Field's conception of the classification problem comes in a discussion of the "ghost of the Tarski hierarchy" quotation from Kripke (1975). Field points out that this quotation comes from an argument directed against the possibility of a language that can express all concepts. Field argues persuasively that Kripke's arguments here do not work. Field goes on to say that what Kripke should have wanted was a theory of truth "that there is no reason to go beyond simply to say what we want to say about truth and related concepts."[26]

There are two sources of potential ambiguity in this version. The first is the involvement of the theorist. Depending on one's other philosophical commitments and aims, this can vary. If one is a Tarskian theorist, for example, one may want to define truth. If one is teaching introductory logic, one may not want to say anything about self-referential truth. If one is aiming at characterizing the truth predicate in natural language, one may want to talk about self-reference and truth without defining the extension of the truth predicate. There does not seem to be a natural, privileged set of things one might want to say about truth independent of the specification of these further goals.

The other source of ambiguity is the phrase "saying what we want about truth and related concepts." There are at least four things this could mean. First, one might just want to use the truth predicate to say that some sentences are true. Second, one might want to state and derive semantical laws. Third, one might want to describe semantically defective sentences as such or otherwise indicate when application of the truth predicate is inappropriate. Fourth, one might want to provide an explicit definition of truth. These are all things that one could reasonably want to say with the truth predicate and related predicates.

There is little hope that Field's theory will allow one to say everything one wants about truth if one wants to say things about truth that involve diagnostic predicates, such as truth-teller-like or liar-like, in the object language. There is nothing in the object language of Field's theory that distinguishes liars and truth-tellers. The two kinds of sentences have the same semantic value through all stages of Field's model construction. Field's theory

[26] Field (2008, 222)

can be modified to distinguish liars from truth-tellers, in the same way that all fixed-point theories can be.[27] It is worth noting, however, that Field's conditional does permit him to define a predicate that is true of all sentences that receive value **n** constantly throughout the model construction: $Tx \leftrightarrow \sim Tx$, where '$\leftrightarrow$' is the biconditional of Field's theory. In the modified form of Field's construction, this predicate will be true of all liar-like sentences, but there is not an analogous predicate for truth-tellers.

Field's conception needs to be fleshed out with an indication of the limits of what he wants to say, or what one should want to say, about truth and related notions. There does not seem to be anything built into the concept of truth that would fix such limits at the outset.

The classification problem has another dimension to it that can be brought out by looking at claims made elsewhere by Field.[28] Field wants to classify the semantic status of all sentences in his object languages using the resources of the object languages. He adopts the following principle.

Status For all claims A, there is some semantic predicate that can be used to classify correctly the semantic status of A.

The predicate P correctly classifying A requires that the predication $P(\ulcorner A \urcorner)$ be true. As a consequence of this together with some features of the semantic status predicates, every defective sentence is classified as defective in some way or other without there being a single, unified concept of semantic defect. Field wants to classify all paradoxical sentences as defective in some way.[29] Rather than a general notion of semantic defectiveness, he uses a definable hierarchy of determinateness operators. There is no level of the hierarchy at which every defective claim is labelled as such, but for every defective claim there is some level at which it is labelled as defective. Field adopts the following $\forall\exists$ principle, as a consequence of Status, above.

[27] Field considers and rejects this option. See Field (2008, 271-274). This rejection is not forced by other commitments. The option Field considers does not provide a predicate true of all and only liars or true of all and only truth-tellers.

[28] Fitch also makes similar claims. See Fitch (1964) and Fitch (1946) for examples. I omit the details of Fitch's theory, which is in many respects similar to that of the least fixed-point strong Kleene theory. There is a detailed reconstruction in Updike (2010).

[29] This phrasing, "defective in some way" is my own, not Field's.

Defect For all defective claims A, there is some semantic predicate that can be used to say that A is defective.

Field adopts Defect while denying that, for a given language, there is any unified concept of semantic defect with which to classify all defective sentences.

Field views the classification problem as a central one for the theory of truth and he thinks that an $\forall\exists$ claim, such as Defect, is sufficient for solving the classification problem. Some philosophers, such as Priest, think that the corresponding $\exists\forall$ version of Defect is necessary to solve the classification problem. Priest's reasons are bound up with his criticisms of hierarchical semantic concepts, which I discussed in §1.7. Does the descriptive project supply definitive evidence in favor of either an $\forall\exists$ version of Defect or an $\exists\forall$ version? As I said in the previous section, Kripke felt pressure to be able to classify all the sentences of the language of his theory. Priest would, I think, cite my statement of Defect as supplying reason enough to adopt the $\exists\forall$ formulation.

2.3.3 Beall

Beall notes that a kind of classification problem has been prevalent in recent work on truth.[30] The problem is to show how a theory of truth can classify the semantic status of all the sentences of the language of the theory itself. Beall describes the exhaustive characterization project (ECP), which answers this problem, as follows.

ECP This project aims to explain "how, if at all, we can truly characterize —specify the 'semantic status' of— all sentences of our language (in our language)."[31]

Beall thinks that the liar paradox makes it hard to see how we can have an exhaustive characterization of the semantic status of all the sentences in the language of our theory as well as a non-trivial theory of truth.[32]

A conceptual question that Beall does not address is why recent work on truth has adopted ECP. We should distinguish two reasons for adopting ECP. One is adopting it out

[30] Beall (2009, 66)

[31] Beall (2009, 66). See Beall (2009, 134-141), Jenkins (2007), Caret and Cotnoir (2008), and Jenkins (2008) for further discussion of ECP.

[32] Fitch also seems to think that ECP is a project that every theory of truth must adopt. See Fitch (1964) for discussion.

of technical interest; can a theory be constructed that successfully executes the project? The other is adopting it because one believes that a semantic theory for natural language should be able to do what ECP says. The former reason corresponds to possible closure and the latter reason, on which I will focus, corresponds to actual closure.

It does seem to be the case that we can classify the semantic status of sentences, such as the liar, in some way. This does not mean that a theory of truth has to formalize the classification as, say, gappy. Liar sentences should be distinguished from contingently false sentences and from explicit contradictions. Similarly, sentences such as truth-tellers should be distinguished from contingently false sentences and contingently true sentences. Liars and truth-tellers seem to behave different semantically, and this is a feature that one can recognize by reflecting on arguments involving those two types of sentences. There is an intuitive distinction of these two types of pathological sentences, and this intuitive classification lends some support to the need for diagnostic predicates and not just semantic value predicates.

There are at least two demands we can make on the way an object language semantic theory handles the status sentences. We could require that the theory entail the status of particular sentences, as in a ground model. This is an excessively large demand. Alternatively, we could require that the theory entails general facts about the statuses, such as semantical laws for them.

The theory of truth is supposed to formalize the logic of truth and semantic features of the truth predicate. There is no reason to require that the theory entails the status of every sentence of an object language. Rather, it should generate the logic and general laws about truth and semantic vocabulary. The language of the theory can contain predicates that could be used to classify all the sentences correctly, but the theory need not entail all of these status attributions.

The next question I want to raise concerns what ECP is demanding. More specifically, the question is what the available semantic statuses of the language might be. The standard semantic values, such as true, false, neither true nor false, and both true and false, all seem reasonable as candidates. In the previous section, I discussed the $\forall\exists$ claim Defect and its $\exists\forall$ variant. ECP only supports the $\forall\exists$ version. ECP would be satisfied if there

were enough predicates to classify the semantically defective sentences without there being a single predicate that does that classification.

There are other status predicates, the diagnostic ones, that describe particular semantic and logical features of sentences. These include truth-teller-like, liar-like, and pathological. There is good reason, which was Kripke's motivation for adopting the descriptive project, to want to use the predicates to characterize the semantic status of our sentences.

While the diagnostic predicates are not traditional semantic statuses, such as the truth predicate, they have a claim to being legitimate semantic statuses. If a sentence exhibits a particular sort of semantic pathology, such as being truth-teller-like, the theory of truth, or the semantic theory of which it is a part, should be able to describe it as such. On Beall's proposal, or any proposal that is based on the minimal fixed-point, the distinction between liar-like and truth-teller-like is collapsed, receiving the same value in all the minimal fixed-points.

There is another diagnostic notion in the vicinity that Beall himself uses in the presentation of his theory, that of being Curry-like. Beall's use of the notion of Curry-like when he is presenting his theory indicates that the diagnostic notion is important for his theory. It is in line with the rest of his view to want to express that classification in the object language theory, but it is not clear whether the theory has the resources to do so.

I will now turn to a discussion of diagnostic predicates.

2.3.4 Diagnosis

In §2.3.1, I distinguished semantic value predicates from diagnostic predicates. Much of the discussion around Semantic Closure has focused on the semantic value predicates. There is a strain of thought that is lost in many of those discussions: a theory of truth should provide us with resources to analyze intuitive distinctions amongst sentences, whether they correspond to semantic values or not. For example, a liar sentence and a truth-teller sentence are both odd, but they are odd in different ways. The theory of truth should provide us with the resources to classify sentences as being odd in the same way as the liar and as odd in the same way as the truth-teller.

Kripke, Visser, and Gupta and Belnap offer ways of analyzing intuitive distinctions of the sort exemplified above.[33] One of the conceptual advances over the Tarskian approach offered by the fixed-point approach of Kripke, Martin, and Woodruff is that the latter not only allows sentences that circularly attribute truth but also provides ways of distinguishing semantic differences in those attributions. Many proposals that allegedly solve the problem of Semantic Closure are unable to account for these distinctions. For example, Field's preferred theory does not accommodate a distinction between truth-teller-like sentences and liar-like ones.[34] Both are simply not determinately true. Field's theory does not offer any reconstruction of the distinction between determinate liar-like sentences and Curry-like ones. Instances of both are simply not determinately true for n iterations of the determinateness operator, symbolically $\sim D^n$, for $n \geq 2$.

The concepts of truth-teller-like and liar-like are, arguably, not pre-theoretic concepts, but they do reflect intuitive, pre-theoretic distinctions. Those concepts are related to truth and semantic notions, so they are the sort of thing that a descriptive theory of truth should accommodate.

There are some diagnostic predicates, such as "gappy" in strong Kleene theories or "just true" in LP theories, that have received a fair amount of attention and discussion. I am not sure that accommodating diagnostic predicates requires having those to be available in the theory. Even if the non-classical approaches to truth have strong reasons to bar those predicates, other diagnostic predicates, such as liar-like and truth-teller-like are not touched by those arguments, because the latter reflect semantic and logical differences whose existence everyone should agree upon.

Diagnostic predicates can give us a clearer picture of what the logic of truth is. They can, additionally, indicate differences in the types of self-reference and the conceptual similarities between self-referential reasoning in different domains. If a theory of truth cannot accommodate some diagnostic predicates, then the theory gives up on a part of the descriptive project, and abandoning that part counts against such a theory of truth.

A theory of truth does not need to contain all the diagnostic predicates in the object

[33] See Kripke (1975, 72-74), Visser (2004, 170), and Gupta and Belnap (1993, 188-189), respectively.

[34] The theory can be modified to use these notions, but, as I will discuss shortly, Field tentatively rejects this suggestion.

language. This is a consequence of the arguments from §1.7 against problems stemming from semantic hierarchies and the arguments to come in §2.4 against problems stemming from certain uses of metalanguages for theories of truth.

There are approaches to truth that claim that everything that can legitimately be said about truth can be said in the particular object language of the approach.[35] Some of the diagnostic predicates, however, resist incorporation into the object language theories. For example, in the strong Kleene fixed-point approach, there is a notion of truth-teller-like that is defined in terms of the class of all fixed-points over a ground model. A sentence is truth-teller-like if and only if there are fixed-points in which that sentence has each of the semantic values **n**, **t**, and **f**. I will now sketch a proof that the strong Kleene fixed-point approach cannot contain its own truth-teller-like predicate in the object language.[36]

Suppose that τ is a predicate interpreted as truth-teller-like and there is a term, a, where $a =$ '$\tau(a) \vee T(a)$'. There are two cases, since either there are the appropriate fixed-points or there are not. Assume that $\tau(a)$ is true. In the minimal fixed-point, $T(a)$ receives the value **n**, so $\tau(a) \vee T(a)$ receives **t**. Then there are no fixed-points extending that in which $\tau(a) \vee T(a)$ is **f**, which contradicts the assumption. Assume that $\tau(a)$ is **f**. Then $\tau(a) \vee T(a)$ receives **n** in the minimal fixed-point, so there is a fixed-point in which $T(a)$ receives **t** and a fixed-point in which $T(a)$ receives **f**. The sentence $\tau(a) \vee T(a)$ is evaluated in each as **t** and **f**, respectively, which means that $\tau(a) \vee T(a)$ is truth-teller-like, contradicting the assumption. Therefore, we conclude that τ cannot be interpreted as truth-teller-like.

The plain strong Kleene fixed-point theory of truth, for example, accommodates the diagnostic predicates in its metalanguages. Predicates such as liar-like and truth-teller-like cannot be added to the object language, which counts against that theory. The strong Kleene fixed-point theory does accommodate the predicates to some degree, so it does not give up on the descriptive project.

Other theories may use different diagnostic predicates or use a different interpretation of some diagnostic predicates. The revision theory, for example, interprets truth-teller-like in terms of stability in different revision sequences. There are more issues related to diagnostic

[35] Horsten and Field say things along these lines, although it is not clear that either must or that either is committed to this idea. See, for example, Horsten (2011, 144-146) and Field (2008, 222).

[36] This proof strategy was suggested to me by Anil Gupta.

predicates to explore, but I will not do so here. I will sum up the conclusions from this section before moving on to discuss philosophical issues surrounding metalanguages.

2.3.5 Conclusions on classification

In this section, I presented three views on the classification of sentences according to semantic status. I argued that Kripke's claim about the need to classify sentences can be motivated by appeal to the descriptive project that motivates his overall approach to truth. After that, I distinguished two kinds of semantic classification: semantic value and diagnostic.

I then turned to Field's claims about classification and showed that his claims were underspecified. On further inspection, his claims are not supported by his theory. I formulated two principles of classification, Status and Defect, and looked at two versions of the latter, both of which receive some support from the motivating ideas of the descriptive project.

In my discussion of Beall's view, I tried to clarify what the exhaustive characterization project is and why one should adopt it. These considerations lend support to diagnostic predicates. I then turned to a discussion of diagnostic predicates and why they should be included in a theory of truth. They present a *prima facie* problem for some theories that are presented as solving the problems of Semantic Closure, as they are not compatible with many such theories, such as some interpretations of the strong Kleene fixed-point theory or Field's modified fixed-point approach.

The upshot of this section is that a theory of truth needs to have available both semantic status and diagnostic predicates. There are not, as far as I can tell, strong considerations requiring there be a single predicate that covers all of the sentences that one would intuitively classify as defective. There do not seem to be strong considerations in favor of requiring that the diagnostic predicates must be available in a single object language. There are, however, reasons for a theory of truth to accommodate diagnostic predicates in some capacity.

Going just by the classification of sentences into semantic value categories, there is a theory of truth that receives a surprising amount of support. The weak Kleene fixed-point theory extended with a gap predicate can classify all the sentences of its language exhaus-

tively.[37] It is not clear what diagnostic predicates this theory can contain, but it can contain all of its semantic value predicates. The theory does not do as well when other criteria, such as validating T-sentences, are used in addition to the classification of sentences. Nonetheless, it does present a consistent, $\exists\forall$ solution to the classification problem.

I will now turn to the metalanguage problem.

2.4 METALANGUAGES

Metalanguages have played an important role in discussions of theories of truth and Semantic Closure. Indeed, in §1.4, I offered the metalanguage problem as a key component in my analysis of the broad problem of Semantic Closure. The metalanguage problem is the problem of providing a theory of truth that has a semantic metalanguage that is a part of the object language. In this section I will examine the motivation for this problem.

Despite the centrality of the metalanguage problem, philosophers do not say much about the philosophical role of metalanguages in theories of truth.[38] For example, Field (2008) claims as goal the construction of languages "that are sufficiently powerful to serve as their own metalanguages."[39] The contrast is a language with a theory of truth that is "indefinitely extensible," that is, "there is a simple recipe whereby given any precise theory of truth, we can automatically construct a more extensive one."[40] Field's goal seems to be to implement a principle defended by Reinhardt, which I will criticize in §2.4.1. Field does not say much more about why we need metalanguages such as the ones he describes, nor does he indicate the problem with theories that comes with recipes for generating more extensive theories.

Concern over the relation between metalanguages for theories of truth of their object languages seems to have its origin in two places, worries over the role of metalanguages in the Tarskian approach and worries over the need to use distinct metalanguages to say certain things with the fixed-point approach.

[37] See Gupta and Martin (1984) for details.

[38] See Visser (2004, 163-164) for a complaint about the confusion and conflation of the role played by metalanguages with the roles played by other notions, including type theory and the use-mention distinction.

[39] Field (2008, 18)

[40] Field (2008, 18)

I discussed the Tarskian view in §2.1 and I discussed Kripke's comment in §2.3.1. In this section, I will focus on philosophers reacting to the work of Tarski and Kripke. I will begin by reconstructing an argument by Reinhardt provided in response to Kripke's comment about the need to ascend to a metalanguage. After I discuss Reinhardt's argument, I will present Priest's criticisms of the distinction between metalanguage and object language, and then I will discuss McGee's argument for the inadequacy of theories of truth that use that distinction. I will now turn to Reinhardt's view.

2.4.1 Reinhardt

Reinhardt suggests a way of drawing the distinction between metalanguage and object language: the distinction between one language and another which contains a truth predicate for the first.[41] Reinhardt's goal is to explain what is needed to provide an interpretation of a formal language. He says that often informal concepts are used to describe the intended interpretation of the formal language. He goes on to say,

> 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 pre-existing informal notion of truth. If we now try to apply the [above way of drawing the distinction between metalanguage and object language], we distinguish between L and L together with truth for L. But truth is *already* a predicate of L! Unless we are prepared to entertain a splitting of 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.[42]

Reinhardt's conclusion is his desideratum for theories of truth, that the object language in question must contain its own metalanguage, in his sense of the terms. Reinhardt's conclusion is that L contains a self-referential truth predicate. Reinhardt's argument is circular, since we assumed at the outset that L contains its own truth predicate.

If Reinhardt's conclusion concerns other ways of drawing the distinction between object language and metalanguage, then Reinhardt is equivocating on the distinction. Reinhardt's argument provides no reason against using a stronger or more expressive metalanguage, in

[41] Reinhardt (1986, 226)
[42] Reinhardt (1986, 227-228)

the more traditional sense of the term. His comment about splitting the notion of truth is a non-sequitur. The appropriate conclusion to draw from Reinhardt's argument is that the object language of a theory of truth should contain a self-referential truth predicate. We have independent motivation for using a self-referential truth predicate.

I will now turn to Priest's criticisms of the object language/metalanguage distinction.

2.4.2 Priest on metalanguages

Priest is a harsh critic of attempts to block paradoxes using the distinction between object language and metalanguage. One of Priest's primary motivations for criticizing the use of distinction in theories of truth is that he wants an $\exists\forall$ semantic theory, which is to say that he thinks there is a single theory that correctly characterizes the semantic status of all the sentences in the language of the theory. Priest says,

> The whole distinction between object theory and metatheory should be abolished, at least in the sense that it is normally understood. (That is, that the metatheory must be a different, and in fact stronger, theory than the object theory.) ... A natural language (or a formal language that models that aspect of its behaviour) *can give its own semantics.*[43]

Priest does not give justification for his claim that a natural language can give its own semantics. I think that he has something like the following argument in mind. There are adequate semantic theories for natural languages. Those theories are in some language or other. If there are parts of the theories not in our language, then we can add the relevant vocabulary and axioms. Thus, the language can give its own semantic theory.

There is an immediate question of why one should think that the theory in our language will be adequate, or even non-trivial, given only that there is an adequate theory in some other language. This is difficult to settle in the abstract, and I think that one can grant this point, for the sake of argument, and still respond to the preceding argument. The response proceeds by addressing two natural questions. First, can natural languages, in fact, give their own semantic theories? Second, if they can, are they all of the $\exists\forall$ sort that Priest thinks they are?

[43](Priest, 2006, 70), emphasis mine.

With respect to the first question, there are not any extant semantic theories for full natural languages, even theories formulated with a paraconsistent logic. This need not lead to skepticism about the whole enterprise of semantics. I will assume that Priest is correct in his claim that natural languages can give their own semantics. If we can give a semantic theory for a natural language in that language, must it be of the $\exists\forall$ kind?

In §1.7, I distinguished a part of Priest's objection to semantic hierarchies: the use of additional predicates in the metalanguage commits one to using those predicates in an extension of the object language. Those predicates must be incorporated into the object language. Priest's reason for this part of his objection is that the metalanguage predicates are required for an account of the object language predicates. Since the object language truth predicate is supposed to reflect or model the real concept of truth, we need to consider the language that has both truth and the additional predicates. Priest would say that any account of truth lacking a language that contains the predicates that the theory generates is inadequate as an account of truth. McGee makes the same claim about the inadequacy of theories of truth in his discussion of metalanguages for theories of truth. I will respond to both in my response to McGee in the next section.

2.4.3 No Richer Metalanguages

McGee is primarily interested in the actual closure question of §1.5. He thinks that a particular form of semantic closure is a requirement on theories of truth, namely that the metalanguage be the same as the object language. He thinks this because otherwise the theory, or rather the semantic theory of which it is a part, will render meaning in natural language utterly mysterious. Non-semantically closed theories of truth make meaning mysterious because they require a richer metalanguage that does not exist. In this section, I will present and criticize McGee's argument.

McGee motivates his argument by saying the following.

> To understand how the English language works, we need an account that encompasses the language as a whole; an infinitely fragmented account will not do. To be sure, we can get about in the world even if our semantic theory is infinitely fragmented, but then again we can get about in the world with no semantic theory at all.[44]

[44] McGee (1991, 80-81)

I have two comments on this. The first is that McGee sometimes sounds as if he thinks that a theory of truth that validates all of the T-sentences would be an adequate semantic theory.[45] If validating the T-sentences is all we require of a semantic theory, then it will not be hard to obtain an adequate semantic theory and theory of truth.

The second comment is that McGee's claim that in order to understand how English works, we need an account of meaning for the whole language is not true. Issues connected to understanding were dealt with in §1.8, so I will not focus on those aspects of McGee's claim. Consider a claim analogous to McGee's but about logic: to understand how modal logic works, we need an account of the whole modal language, and looking at fragments, such as the propositional fragment or the positive propositional fragment, will not do. This is not correct, since investigating the parts of the broader phenomenon help us to get clearer on the logic of those parts and their relations to each other. A full axiomatization for the propositional and quantificational parts of the language would tell us more, but it would not undermine the contributions of investigating the fragments.

McGee (1991) presents a thesis that he uses to motivate the metalanguage problem.

> The dominant opinion has it that the liar antimony proves that it is never possible to develop a successful theory of truth for a language within the language itself; instead, one must develop the theory of truth for a language $\mathscr{L}$ within a metalanguage that is richer than $\mathscr{L}$ in expressive power. This implies that, *since we have no metalanguage richer than English*, we cannot develop a theory of truth for English, or for any natural language.[46]

McGee thinks that we must use a theory of truth-for-$\mathscr{L}$ that can be developed in $\mathscr{L}$. His reason is encapsulated in the following principle.

No Richer Metalanguages There is no language that is essentially richer than any distinct natural language.

By this principle, there is no metalanguage for, say, English that is more expressive than English.

McGee uses the No Richer Metalanguages (NRM) principle as a premiss in arguments against many theories of truth. McGee argues that the revision theory is not a theory of truth for English because it requires a metalanguage richer than its object language. NRM

[45] See, for example, McGee (1991, 78).
[46] McGee (1991, ix), emphasis added.

says there is no such metalanguage for English, so the revision theory cannot take English as an object language. Therefore, the revision theory is inadequate as a theory of truth for natural language.[47]

Using NRM, we can construct arguments against many theories of truth.

1. There is no language that is essentially richer than any natural language. (NRM)
2. Developing theory $\mathscr{T}$ as a theory of truth-for-$\mathscr{L}$ requires a metalanguage essentially stronger than $\mathscr{L}$.
3. Therefore, theory $\mathscr{T}$ is inadequate as a theory of truth for natural languages.

In this argument, NRM is the important premiss. There are two questions that we should ask about NRM. What is meant by "essentially richer metalanguage?" Is NRM true? I will take these in turn.

There is a determinate meaning to the phrase "essentially richer metalanguage" in the case of a Tarski's theory of truth.[48] The truth predicate for a language is restricted and the addition of that truth predicate to the object language results in inconsistency. The metalanguage has additional resources that the object language does not and cannot contain. In the case of the revision theory, the meaning is less clear. The object language may not contain a categoricalness predicate, but one can be added to it without contradiction.

I see two primary ways to argue for NRM. The first way uses Extensibility, from §2.1.2. For any putatively richer language, add the vocabulary constituting the additional richness to the impoverished natural language. The richer language is then not richer after all. The second way is a general argument against attempts to falsify NRM. If someone offers a counterexample to NRM, it will be in some language. So, the counterexample will be in a natural language, the one the objector is using, which may even be English. It will, then, not be a counterexample after all.

There are two responses I want to give to the NRM argument. The first assumes that Extensibility is used to argue for NRM. We need to make a distinction to clarify what language is under consideration. There are two options, the natural language at a particular

[47] See McGee (1991, 147 ff.) for McGee's version of this argument.

[48] This is a controversial claim, but it is, I think, the best case for McGee. See DeVidi and Solomon (1999) and Ray (2005) for discussion.

time slice or the final product of all possible extensions, one that is closed under further additions of vocabulary. If the latter, then it not surprising that none of the current methods of constructing theories of truth will work since they require a specification of the resources in the language. So, in that case the argument is sound but not interesting.[49] The second option, using time slices of languages, makes the first premiss false. There is not any reason to think that later time developments of a language will be no richer, in the appropriate sense, than the stage under consideration. Indeed, if some process of reflection on the semantic features of a language leads to the introduction of new semantic vocabulary, as suggested by some theories of truth, then there will be outright counterexamples to the first premiss.

The descriptive project does not require making sense of the result of all possible extensions. In §2.1.2, I argued against the $\exists\forall$ interpretation of Extensibility. In addition to that, we do not have any access to the result of all possible extensions to determine what sorts of arguments come out valid using its logic. The descriptive project can motivate theories of truth for our language as it stands now as well as under extensions, but that does not entail that the project motivates theories of truth for the language resulting from all possible extensions.

The second response I have to the NRM argument is that it is invalid if we admit that partial accounts or descriptions of the semantics of natural languages can be adequate as semantic theories.[50] Adequacy, here, consists in providing an account of both the ways in which semantic values of parts of sentences determine the semantic values of whole sentences and the logical relations among sentences. Suppose that we are working with a theory that provides such a description, say a revision theory. By hypothesis, the theory cannot provide the whole description of the semantics of the language, but there are systematic ways of developing revision theories of semantical concepts besides truth.[51] Individually, these theories provide a partial description of the semantics of the language, which satisfies the partial account desideratum. This means, however, that such a theory of truth is adequate for the semantics of natural language, contradicting the conclusion of the NRM argument.

[49] This line was suggested to me by James Shaw. It is at root the same as a response in Gupta (1997).

[50] I think that the position adopted in McGee (1997) makes such an admission.

[51] See Gupta and Belnap (1993), particularly chapter 7, for a discussion of revision approaches to other semantic concepts.

My conclusion is that the NRM argument cannot be used to dismiss theories of truth. According to plausible views of semantics, the argument is invalid. Even if those views are rejected, I argued that the NRM argument is unsound and not supported by the descriptive project. We should reject NRM and, with it, Priest's concerns over metalanguages.

2.4.4 Conclusions on metalanguages

The work done in this section was primarily negative. In this section, I presented three different views on the philosophical relationship between metalanguages and object languages in adequate theories of truth. I argued that Reinhardt's argument was circular. After this, I examined an argument by Priest and found it to depend on a principle advocated by McGee. I presented McGee's argument for that principle, and indicated how it depends on acceptance of Extensibility ($\exists\forall$) from §2.1.2. Analyzing McGee's argument with my distinctions in place shows that his argument is either sound, but uninteresting, or it is invalid.

I will now turn to a topic not discussed at length by any of the philosophers at which I have looked.

2.5 LAWS

In §1.6, I listed several things that might be needed in order to provide a semantic theory for a formal language and then listed recursion clauses for computing truth values of complex sentences. If these recursion clauses are rewritten with a truth predicate, then we have the *semantical laws for truth*. None of the philosophers discussed defend the semantical laws as components of Semantic Closure.[52] I will argue that the semantical laws figure in a necessary condition on theories that adequately address Semantic Closure. The purpose of this section is to explain what the semantical laws are and to argue that they are necessary conditions on solutions to the broad problem of Semantic Closure.

The semantical laws for truth are the laws that govern how truth interacts with the

[52] Field (1999, 534-536) cites the semantical laws as important for the function of truth.

logical connectives. If there are other central semantic predicates, then there will be other semantical laws for those predicates. The semantical laws for truth include the universal closures of the following, where A and B range over sentences of the language in question.[53]

- $T(\ulcorner A \,\&\, B \urcorner) \Leftrightarrow T(\ulcorner A \urcorner) \,\&\, T(\ulcorner B \urcorner)$
- $T(\ulcorner A \vee B \urcorner) \Leftrightarrow T(\ulcorner A \urcorner) \vee T(\ulcorner B \urcorner)$
- $T(\ulcorner \sim A \urcorner) \Leftrightarrow \sim T(\ulcorner A \urcorner)$
- $T(\ulcorner \forall x A x \urcorner) \Leftrightarrow \forall t T(\ulcorner A[x/t] \urcorner)$
- $T(\ulcorner \exists x A x \urcorner) \Leftrightarrow \exists t T(\ulcorner A[x/t] \urcorner)$

In the quantifier laws, t ranges over closed terms. The right-to left direction of the universal quantifier law may fail if there are not enough terms in the language to name all objects.[54]

As I said at the start of this section, there are close correspondences between the semantical laws and the recursion clauses from the definition of valuations from §1.6. The recursion clauses, however, involved terms from two languages, the object language over which the valuation is defined and the metalanguage in which the semantical computations are being carried out. The semantical laws, by contrast, are in a single language, the language of the object theory with the truth predicate. They express an analog of the recursive computations of the recursion clauses, saying how to compute the truth of a complex sentence on the basis of its components.

There is a correspondence between the semantical laws and the recursion clauses, and this indicates the importance of the semantical laws for the broad problem of Semantic Closure. The recursion clauses are a key feature of a compositional semantic theory. They spell out how the truth value of complex sentences is computed from the truth value of simpler sentences. In order for a theory of truth to play the role of a semantic theory, it needs to state how the truth of complex sentences relates to the truth of simpler sentences, and similarly for other semantic predicates. One of the motivating intuitions behind the broad problem of Semantic Closure is that a theory of truth should have a close connection

[53] According to the conversion of §1.2, '$\Leftrightarrow$' is a generic biconditional. Different theories will plug in different biconditionals for their versions of the semantical laws.

[54] Nuel Belnap has pointed out that the quantifier laws force a substitutional interpretation of the quantifiers. This leads to an objection, namely that we frequently do not have names for everything about which we want to talk, even if there are only finitely many such things. This may be reason to reject one or both directions of the quantifier laws. The objection merits further discussion, but I will put it aside.

with the broader semantic theory. Much may be involved in the connection, but some analog of the recursion clauses in the language of the theory will be needed.

There are three things I want to point out concerning the semantical laws. First is a difference between the semantical laws and the base recursion clauses. The base recursion clauses say how to determine whether an atomic sentence is true. The semantical laws above do not address this. The relation between truth and atomic sentences is fixed by the T-sentences. The semantical laws deal with computing the truth of a complex sentence on the basis of the truth of its components.

Second, the semantical laws do not specify whether iterated truth predicates are equivalent. Define T^0A as A and $T^{n+1}A$ as $T(\ulcorner T^nA \urcorner)$. The *T-step* sentences are biconditionals of the form

$$T^nA \Leftrightarrow T^mA,$$

for $n \neq m$. These say that iterations of the truth predicate are equivalent.

There are some special cases of T-step that are worth pointing out. Let us say that the (n, m) T-step biconditionals are those whose left-hand side is T^n and whose right-hand side is T^m. Among the (1,0) T-step biconditionals are instances of the T-sentences for truth-free atomic sentences.[55] The semantic ideas that motivate the semantical laws motivate these as the base cases of the recursion clauses. The set of the $(n+1, n)$ T-step biconditionals, for each n, gives the set of T-sentences for the whole language. My point in highlighting these is that the semantic ideas motivate some, but not all, of the T-step biconditionals.

The revision theory does not validate any of the T-step schemes.[56] The strong Kleene fixed-point theory does not validate the T-step schemes that use the material biconditional,

$$T^nA \Leftrightarrow T^mA,$$

but it does validate the T-step schemes if the biconditional is the following metalanguage material equivalence,

$$M + (\mathscr{T}, \mathscr{F}) \models T^nA \text{ iff } M + (\mathscr{T}, \mathscr{F}) \models T^mA.$$

[55] Also among the (1,0) T-step biconditionals are sentences like $T(\ulcorner \sim T(\ulcorner p \urcorner) \urcorner) \Leftrightarrow \sim T(\ulcorner p \urcorner)$ and $T(\ulcorner p \,\&\, q \urcorner) \Leftrightarrow p \,\&\, q$.

[56] See Gupta and Belnap (1993, 221) theorem 6C.5 for details.

Field's theory validates the T-step scheme that uses the Field conditional. The augmented revision theory of the next chapter validates the $(n+1, n)$ T-step scheme that uses the biconditional proposed there.

Third, the semantical laws use a biconditional connective. This need not be the material biconditional. Other biconditional connectives can be used, as appropriate to the theory. In some theories of truth, there is one biconditional connective used for the T-sentences and another used for the semantical laws. For example, the revision theory of truth uses a definitional connective for the T-sentences and a material biconditional for the semantical laws. It may seem more natural to use the same connective for both, but this is not forced on us. There are ways of understanding the T-sentences that make acceptable the use of biconditionals different from those in the semantical laws. In the revision theory, for example, the T-sentences define the notion of truth while the semantical laws are consequences of that definition.

The semantical laws I have listed all involve logical vocabulary that is generally thought of as extensional, or truth-functional. There may be other logical vocabulary in the language, some of it intensional, such as the '$\Box$' of logical or metaphysical necessity. If there is other logical vocabulary, then there will need to be additional semantical laws. The law for the box of metaphysical necessity would be the universal closure of the following.[57]

- $T(\ulcorner \Box A \urcorner) \Leftrightarrow \Box T(\ulcorner A \urcorner)$

Not all philosophers accept that the semantical laws should be validated by their theories, although many insist that truth commute at least with the truth-functional connectives. Priest, for example, thinks that the left to right direction of the negation law does not hold. This leads to a gap between the metalinguistic semantic evaluation of complex sentences using his model-theoretic semantics and the object language semantic evaluation using the semantical laws.

The truth predicate should commute with the extensional connectives, but not only with them. There are intensional connectives, such as necessity operators, with which the truth predicate should commute as well. There are some intensional connectives with which it

[57] There are some subtle issues that I am eliding regarding the distinction between propositions and sentences. See Gupta (1978) for discussion. I owe this point to Nuel Belnap.

need not commute. If we took 'said "…"' to be a logical connective, then truth need not to commute with it. If it is true that Ana said "p", it need not be the case that Ana said "p is true."

Let the *semantical law* for truth for a connective $\bigcirc$ be a biconditional of the following form, where the biconditional is the appropriate one for a given theory.

$$T(\ulcorner\bigcirc(A_1, \ldots, A_n)\urcorner) \Leftrightarrow \bigcirc(T(\ulcorner A_1\urcorner), \ldots, T(\ulcorner A_n\urcorner))$$

The following principle, which connects semantical laws and adequately semantically closed theories of truth, holds.

Semantical Laws A theory of truth is adequately semantically closed only if it validates the semantical laws for truth for all of its logical vocabulary.

Validation of the semantical laws is a necessary, but not sufficient, condition on theories of truth to count as a semantically closed.

For some classical languages, validating the semantical laws will result in ω-inconsistency.[58] It is not clear from the considerations above what the appropriate response to ω-inconsistency is. When comparing theories of truth, ω-inconsistency is one factor for evaluation, but it seems to be distinct from the broad problem of Semantic Closure.

2.6 CONCLUSIONS

In this chapter, I analyzed the remaining aspects of the broad problem of Semantic Closure. Before proceeding to the next chapter, I will take stock of my positive conclusions so far.

In §2.1, I argued that the descriptive project supported one form of Extensibility, namely Extensibility $\forall\exists$. I used this, in turn, to argue for Logic Neutrality as a requirement on theories of truth. Any theory that aims to be descriptive should work with base languages that contain any of a range of logical and semantic resources.

[58] See McGee (1985).

In §2.2, I distinguished different closure conditions on theories of truth. I argued that Satisfaction Readiness and Generalized Semantic Closure were supported by appeal to the descriptive project.

The primary positive result of §2.3 was the distinction between semantic value and diagnostic classification. I argued that the same motivations for adopting a theory of self-referential truth, namely the pressures of a descriptive project, provide reason for requiring the theory to accommodate some diagnostic predicates. These predicates capture intuitive distinctions as well as providing more information about the semantic and logical behavior of self-referential sentences. Many theories, primarily fixed-point theories including Field's and a version of Kripke's, fail to accommodate them.

In §2.5, I argued that a theory of truth adequate to be a semantic theory should entail the semantical laws for truth. Further, it should entail these laws not just for the extensional logical vocabulary, but any vocabulary regarded as logical.

The condition of Generalized Semantic Closure requires a theory of truth to validate all of the T-sentences. Validating the semantical laws requires validating universal generalizations of certain biconditionals. These two conditions, in particular, require there to be appropriate biconditionals in the language. This raises the question of what sort of biconditional is needed.

In recent work on truth, there have been attempts to motivate different conditionals on the basis of the needs of a theory of truth. For example, Beall (2009) motivates a kind of relevance conditional on the basis of the need for a conditional that obeys modus ponens but not *ex falso*. Field (2008) introduces a new conditional for several purposes, including validating the T-sentences.[59]

In the next chapter, I will provide some background on conditionals in theories of truth and present a new conditional that is defined for the revision theory. I will argue that it can be used to satisfy the conditions for which I have argued in this chapter. In particular, I will show how the new conditional helps represent the T-sentences, its relation to the semantical laws, and how it supports the logic neutrality requirement.

[59] I will return to Field's theory and his conditional in chapter 6.

3.0 CONDITIONALS AND REVISION THEORY

In the first two chapters, I analyzed different parts of the broad problem of Semantic Closure. My analysis showed that conditionals are central for the adequacy of a theory of truth. There are two primary roles that conditionals play. The first is the expression of *semantical laws*, quantified biconditionals indicating how the truth of a logically complex sentence depends on the truth of simpler sentences, such as $\forall A(T(\ulcorner \sim\!\dot{A} \urcorner) \Leftrightarrow \sim T(\ulcorner A \urcorner))$.[1] The semantical laws should be valid according to the theory. The second is the expression of conditions of adequacy on semantic vocabulary. For theories of truth, these conditions are primarily the T-sentences, sentences of the form $T(\ulcorner A \urcorner) \Leftrightarrow A$. If a satisfaction or "true-of" predicate is used by the theory, then appropriate conditions for those predicates should be expressible and valid.

An inspection of the examples of semantic laws and conditions of adequacy that I have given so far seems to indicate that *biconditionals*, rather than conditionals, are the primary logical notion for my purposes. I will focus on conditionals for two reasons. First, biconditionals are definable from conditionals. Second, there may be conditions of adequacy that are properly expressed with conditionals rather than biconditionals. For example, if there is some semantic predicate that is governed by a law along the lines of the T axiom of modal logic, $F(\ulcorner A \urcorner) \Rightarrow A$, that law will be properly expressed with a conditional rather than a biconditional.

There is a question of what sort of conditional an adequate theory of truth should have. In this chapter, I will propose one kind of conditional and argue that it fulfills its job

[1] In this chapter, I will use '$\Rightarrow$' and '$\Leftrightarrow$' as a generic conditional and a generic biconditional. I will use '$\supset$' and '$\equiv$' for material conditionals and material biconditionals, letting the context determine whether they governed by classical logic or something else. I will use '$\rightarrow$', '$\leftarrow$', and '$\leftrightarrow$' for the conditionals and the biconditional introduced in this chapter.

well.[2] In the final chapter, I will examine a recent proposal by Field and argue that it does not perform well. I will begin by presenting some background on recent approaches to conditionals in theories of truth (§6.1). Following this, I present my proposal (§3.2.1, §3.2.2), explain some of its features (§3.2.3), and respond to several objections to it (§3.4). The full formal definitions and proofs for theorems are presented in the next chapter. The conditionals I propose motivate a particular modal logic, which I will explore in the next chapter as well (§5).

3.1 BACKGROUND

A recent strategy for developing theories of truth is to add a new conditional to the underlying logic. Proponents of this strategy include Field, Yablo, Beall, and Priest.[3] The primary reasons for adding a conditional are to codify or carry out ordinary reasoning and to guarantee that the theory has some desirable feature.

A theory of truth should be able to reconstruct much of ordinary reasoning.[4] Let us distinguish two things this could mean. One is that intuitive laws and ordinary reasoning patterns should be valid with the new conditional. The other thing is that the conditional reduces to the old conditional in unproblematic cases. In Field's theory, for example, there is a special conditional distinct from the strong Kleene material conditional. In cases when the antecedent and consequent both obey excluded middle, the special conditional is equivalent to the strong Kleene material conditional, which in turn is equivalent to the classical material conditional.

The second reason for adding a conditional is to guarantee that a theory has some desirable feature specific to the area of theories of truth. The primary motivation is often

[2]This proposal was developed in joint work with Anil Gupta.

[3]See Field (2008), Yablo (2003), Beall (2009), and Priest (2006), respectively. Brady is not included in this list, because his work, such as Brady (2006), first identifies the basic logic, including a conditional, and then applies it to the theory of truth and to set theory. I hope to make a more comprehensive comparison in future work.

[4]There are various questions about what constitutes ordinary reasoning. I am going to bracket these questions and assume that ordinary reasoning is roughly classical reasoning, so the material conditional does codify ordinary reasoning. For criticisms of the adequacy of the material conditional for codifying reasoning, see Anderson and Belnap (1975).

securing the validity of the T-sentences.

In the literature on truth, there is little discussion of the desiderata for adding conditionals to the logic of a theory of truth. None of the conditionals are motivated by appeal to the sorts of general considerations for which I have argued in the first two chapters.

Field thinks that we need a conditional that will validate ordinary reasoning. This includes validating principles such as the substitution of equivalents. Beall thinks that a theory should have a conditional that obeys *modus ponens*. Priest thinks a theory must validate all of the T-sentences in addition to codifying reasoning.[5]

Field says the following of the strong Kleene material conditional.

> But while [the material conditional] does a passable job as a conditional in the presence of excluded middle, it is totally inadequate as a conditional without excluded middle: with $\supset$ as one's candidate for $\rightarrow$, one wouldn't even get such elementary laws of the conditional as $[A \Rightarrow A]$, $[A \Rightarrow (A \vee B)]$, or the inference from $[A \Rightarrow B$ to $(C \Rightarrow A) \Rightarrow (C \Rightarrow B)]$. ... The lack of a conditional (and also of a biconditional) cripples ordinary reasoning.[6]

In assessing the features of his new conditional, Field says that his conditional "enables us to come much closer to carrying out ordinary reasoning" than the strong Kleene material conditional does.[7] Enabling ordinary reasoning is one of Field's primary reasons for adding a new conditional to the strong Kleene fixed-point theory of truth.

Beall introduces a new conditional so that his theory of truth will have a conditional that obeys *modus ponens*. Beall endorses a fixed-point approach based on LP, whose connectives are those of strong Kleene with the third value, $\mathbf{b}$, designated.[8] *Modus ponens* is invalid for the material conditional in LP. To see this, note that if the antecedent and consequent of a material conditional have the semantic values $\mathbf{b}$ and $\mathbf{f}$, respectively, then the material conditional will have the value $\mathbf{b}$.[9] The inference

$$A, A \supset B \models B$$

[5] See Priest (2006, 55-56). Priest thinks that the reasoning at issue is not captured by the classical material conditional. For a discussion, see Priest (2006, Ch. 6).

[6] Field (2008, 73). Field uses '$\rightarrow$' where I have put '$\Rightarrow$'.

[7] Field (2008, 276). By "ordinary reasoning" here, Field means classical reasoning.

[8] This definition is elaborated upon in 1.2. A more comprehensive introduction to LP can be found in a number of places, including Priest (2006) and Beall (2009).

[9] In symbols: $\sim\mathbf{b} \vee \mathbf{f} = \mathbf{b} \vee \mathbf{f} = \mathbf{b}$

will be invalidated by interpreting A as **b** and B as **f**. Beall seems to want to preserve some ordinary reasoning with his conditional. His conditional, however, does not reduce to the classical material conditional conditional in non-pathological contexts. He views this as a feature of his conditional, presumably because he disagrees with Field over what constitutes ordinary reasoning.[10]

Both Field and Beall motivate the addition of a new conditional by pointing to features of ordinary reasoning that they want to preserve. Both endorse fixed-point theories that weaken the logic, strong Kleene and LP, respectively. Indeed, Feferman says that "nothing like sustained ordinary reasoning can be carried on in [strong Kleene logic]."[11] A new conditional is added to the logic in order to make up for the weakness of the three-valued scheme.

In this chapter, I will work with the classical revision theory, so there is no need to change the scheme. There is, consequently, no change to the interpretation of the material conditional. The need to augment the logic with a new conditional, in order to make up for the deficiencies of the material conditional, does not arise for the revision theory. There are, however, other reasons to add a new conditional, such as expressing valid T-sentences. I will now turn to new conditionals for the revision theory.

3.2 CONNECTIVES

In this section, I begin by defining new conditionals to add to the revision theory (§3.2.1). This is followed by a technical simplification (§3.2.2), which is interesting in its own right. Then I move to discussing general logical features of these connectives (§3.2.3).

3.2.1 Conditionals

In the previous chapter, I argued that theories of truth should exhibit a kind of *logic neutrality*. That is, the theory should be compatible with base languages that contain a wide

[10] See Beall (2009, 119) for discussion.
[11] Feferman (1984, 95). The whole quotation is emphasized in the original.

array of logical resources. In the previous chapters, I looked primarily at fixed-point approaches and revision-theoretic approaches. A major problem with the fixed-point theories is that they do not work with languages with non-monotonic operators. Because fixed-point theories have this problem, I will focus on revision theory. Revision theory does not restrict the logical resources of the base language, so it does better with respect to logic neutrality than the fixed-point approaches. The classical revision theory based on $S^{\#}$ validates all of the semantical laws as well.

The revision theory has its own problems. It does not validate all of the T-sentences that use the material conditional. In fact, some T-sentences turn out contravalid, such as the T-sentence for a liar sentence.

The kind of T-sentence that the revision theory endorses uses a definitional biconditional. The problem is that this biconditional is not the material biconditional and the standard revision theory does not have any way to express this biconditional. The problem of expressing the definitional form of the T-sentences can be solved by adding two new conditionals to the standard revision theory. Before presenting the conditionals, I will need to explain an alteration to the general set up of the revision theory.

The two conditionals to be added are the step-down conditional, '$\rightarrow$', and step-up conditional, '$\leftarrow$'. The intuitive idea behind these conditionals is that they act like cross-stage material conditionals. In the standard revision theory, the material conditional is evaluated with respect to a single stage in the revision process. The semantic value of the material conditional $A \supset B$ at a stage n is determined by the values of A and B at stage n. The semantic value of the step-down conditional $A \rightarrow B$ at stage $n+1$ is determined, in contrast, by the values of A at stage $n+1$ and B at stage n. More precisely, '$A \rightarrow B$' is true at stage $k+1$ if and only if, if 'A' is true at stage $k+1$, then 'B' is true at stage k. Similarly, '$B \leftarrow A$' is true at stage $k+1$ if and only if, if 'A' is true at stage k, then 'B' is true at stage $k+1$.

To extend this idea from truth of a sentence to satisfaction of a formula, we need to change some of the basic definitions from revision theory. We present this material in full detail in the next chapter, so rather than repeat that material here, we will just sketch some of the ideas. We redefine *hypotheses*, relative to a ground model M, to be certain subsets $\mathscr{F} \times \mathscr{V}_M$, where $\mathscr{F}$ is the set of formulas containing no constants and $\mathscr{V}_M$ is the set of

assignments to variables.[12] Assuming $A, B \in \mathscr{F}$, the following clauses govern the semantics of the step conditionals.[13]

$$M + h, v \models A \to B \text{ iff if } M + h, v \models A, \text{ then } \langle B, v \rangle \in h$$

$$M + h, v \models B \leftarrow A \text{ iff if } \langle A, v \rangle \in h, \text{ then } M + h, v \models B$$

Intuitively, $A \to B$ is satisfied at $M + h, v$ just in case either A is not satisfied at the current stage or h "thinks" B is satisfied by v. After revision, h "will think" B was satisfied by v just in case B was satisfied by v prior to revision.

The step biconditional, $A \leftrightarrow B$, is defined as $(A \to B)\,\&\,(A \leftarrow B)$. At this point, the reader may be surprised that the biconditional uses two different conditional connectives. In §3.4, I will argue that this is actually more natural than it may initially seem once it is situated in the broader framework of the revision theory.

I will note that circular definitions in the expanded framework are permitted to contain step conditionals in their *definientia*. This has consequences for circular definitions, particularly for finite definitions, a topic which I will discuss in detail in chapter 4.

A calculus, C_0, for the revision theory based on the semantical system S_0 was presented in Gupta and Belnap (1993). We can augment the calculus C_0 with rules for these two conditionals.

$\begin{array}{l} \underline{A^{k+1}} \\ \vdots \\ B^k \end{array}$	hyp		$\begin{array}{l} A \to B^{k+1} \\ \vdots \\ A^{k+1} \end{array}$	
$A \to B^{k+1}$	$\to$I		B^k	$\to$E

[12] I will usually drop the subscript on $\mathscr{V}_M$.

[13] The restriction to $\mathscr{F}$ is not necessary once the full definitions are in place.

$$\begin{array}{|l|l|l|l|}
\hline
\begin{array}{|l} \underline{A^k} \\ \vdots \\ B^{k+1} \end{array} & \text{hyp} & B \leftarrow A^{k+1} & \\
B \leftarrow A^{k+1} & \leftarrow\text{I} & B^{k+1} & \leftarrow\text{E} \\
\hline
\end{array}$$

The rules for the step-conditionals are simply the rules for the material conditional with different indices on the antecedent and conclusion formulas.

I started this chapter by highlighting two roles for conditionals in theories of truth: expressing semantical laws and expressing T-sentences. The step conditionals allow for the expression of the T-sentences. The resulting T-sentences are valid as well. Using the rules presented above, the T-sentences are derivable. In fact, we can get a more general result, namely that for any definition in a set of definitions $\mathscr{D}$, the definition is derivable using the rules. Suppose $Gt =_{Df} A_G(t)$ is a definition in $\mathscr{D}$.

$$\begin{array}{llll|llll}
\hline
1 & \begin{array}{|l} \underline{A_G(t)^n} \\ Gt^{n+1} \end{array} & \text{hyp} & & 1 & \begin{array}{|l} \underline{Gt^{n+1}} \\ A_G(t)^n \end{array} & \text{hyp} & \\
2 & & \text{DefI} & & 2 & & \text{DefI} & \\
3 & Gt \leftarrow A_G(t)^{n+1} & \leftarrow\text{I} & & 3 & Gt \rightarrow A_G(t)^{n+1} & \rightarrow\text{I} & \\
\hline
\end{array}$$

These proofs together give a categorical proof of $Gt \leftrightarrow A_G(t)^{n+1}$. Since the proof is categorical, it is a proof of that biconditional at every index. When the set $\mathscr{D}$ of definitions is the set of T-sentences, then all of the T-sentences are derivable.

The step conditionals can be used to solve one of the problems with the revision theory from earlier chapters. The T-sentences with the material biconditional are not all valid but the T-sentences with the step biconditional are valid.

One step conditional can be defined from the other. The step-up conditional can be defined from the step-down conditional in the following way. $B \leftarrow A$ is defined as $(\top \rightarrow A) \supset B$. The definition of the step-down conditional in terms of the step-up conditional is similar.

If the sentences A and B are $\mathscr{D}$-free, which is to say they contain no circularly defined predicate or step conditionals, then $A \rightarrow B$ is materially equivalent to $A \supset B$.[14]

1	$A \rightarrow B^{k+1}$	hyp	1	$A \supset B^{k+1}$	hyp
2	$\quad A^{k+1}$	hyp	2	$\quad A^{k+1}$	hyp
3	$\quad B^{k}$	$\rightarrow$E	3	$\quad B^{k+1}$	$\supset$E
4	$\quad B^{k+1}$	IS	4	$\quad B^{k}$	IS
5	$A \supset B^{k+1}$	$\supset$I	5	$A \rightarrow B^{k+1}$	$\rightarrow$I

The special logical character of the step-down conditional is only evident in the presence of circular definitions. As can be seen from the preceding proofs, it is only the consequent that needs to be $\mathscr{D}$-free for the material equivalence of the step-down and material conditionals. To obtain the material equivalence of the step biconditional and the material biconditional, we need both antecedent and consequent to be $\mathscr{D}$-free. Otherwise, the necessary index shift steps of the proofs would be unavailable.

I will now turn to a useful conceptual simplification of the step-down conditional.

3.2.2 Box

The preceding strategy of adding new conditionals may appear slightly complicated. There is a simpler way to proceed. Rather than add two new conditionals, we can add a single operator, $\Box$. The $\Box$ operator is a previous stage operator, and we say that $\Box A$ is true at stage $k+1$ if and only if A is true at stage k The $\Box$ operator can be defined as follows.

$$M + h, v \models \Box A \text{ iff } \langle A, v \rangle \in h$$

[14] If A and B contain no *definienda*, but possibly do contain step conditionals, then $A \supset B$ and $A \rightarrow B$ are materially equivalent. The proof is more complex. There are cases in which the material equivalence holds even if A and B contain *definienda*.

The $\Box$ has rules of inference that are similar to the truth predicate.

$$\begin{array}{l|l} A^k & \Box A^{k+1} \\ \Box A^{k+1} \quad \Box\text{I} & A^k \quad \Box\text{E} \end{array}$$

The step conditionals can be defined from the $\Box$ operator in the following way.

- $A \rightarrow B =_{Df} A \supset \Box B$
- $B \leftarrow A =_{Df} \Box A \supset B$

The $\Box$ operator can be defined from the step-down conditional as $\Box A =_{Df} \top \rightarrow A$. The $\Box$ operator and the step conditionals can be defined from satisfaction, if there is a satisfaction predicate in the language. I will work exclusively with the box in the next chapter, because the box is conceptually simpler and easier to use. The box motivates a particular modal logic, which has not, as far as I know, been explored much. I will study it in chapter 5.

3.2.3 Features

I propose to augment the standard revision theory with the box or the step-down conditional. There is then a question of what sort of logic these connectives have. In this section I will discuss this issue.

The box commutes with all logical connectives and quantifiers. The conditionals are a bit more interesting. I will begin by presenting some rules and theorems that fail for the step-down conditional and the step biconditional. Perhaps the most surprising one, at first blush, is that $A \rightarrow A$ fails. This should be clear when we consider a circularly defined predicate G. If we have Gt^{i+1}, we need not have Gt^i, although $A(t, G)^i$ will be derivable by the DefE rule. For $\mathscr{D}$-free formulas A, $A \rightarrow A$ will be derivable.

The question is whether the failure of $A \rightarrow A$ is a large flaw with the conditional. I think it is not. The step-down conditional, or rather the step biconditional, reflects a certain connection between the antecedent and consequent, namely the connection of *definitional equivalence*. The view of Gupta and Belnap (1993) makes a distinction between definitional equivalence and material equivalence. This is to allow a wider class of circular definitions

to be treated in a non-trivial manner. In general, definitional equivalence is distinct from material equivalence, and the former does not entail the latter. For example, the definition $Gx =_{Df} \sim Gx$ is pathological but non-trivial on the revision theory. If definitional equivalence were material equivalence, then the result would be a contradiction. The failure of reflexivity is one indication of the distinction between the two kinds of equivalence.

The step conditionals are playing a particular role in the theory. They reflect the connection definitional equivalence in some of their uses. The step conditionals are not meant to take over the job of primary conditional in the logic. They are two of three conditionals in the logic, the other being the standard material conditional of classical logic. Other approaches to truth do not make a distinction between definitional and material equivalence, so it is less natural for those theories to use multiple conditionals in the way the revision theory does.

The revision theory can treat the definition $Gx =_{Df} \sim Gx$ consistently. If reflexivity is imposed, then the distinction between definitional and material equivalence collapses. Here

is the proof, given a definition $A =_{Df} B$.[15]

1	$A \leftrightarrow A$	
2	$A \leftrightarrow B$	
3	$\quad A^k$	hyp
4	$\quad A^{k+1}$	$\leftarrow$E 1, 3
5	$\quad B^k$	$\rightarrow$E 2, 4
6	$A \supset B^k$	$\supset$I 3-5
7	$\quad B^k$	hyp
8	$\quad A^{k+1}$	$\leftarrow$E 2, 7
9	$\quad A^k$	$\rightarrow$E 1, 8
10	$B \supset A^k$	$\supset$I 7-9

If reflexivity is imposed on definitional equivalence, then a step biconditional entails its material biconditional counterpart. This fact combined with the definition $A =_{Df} \sim A$ results in the classically inconsistent $A \equiv \sim A$.

Another possibly surprising feature of the step biconditional is that it is not symmetric. Symmetry should fail on the interpretation that the step biconditional is given, because the interpretation is that of definitional equivalence. Symmetry would mean that the *definiens* and the *definiendum* could be exchanged freely in the definition. The failure of symmetry for the biconditional is not, I think, a flaw.

[15] In the following Fitch proofs, I use some sentences without indices. This indicates that they are available at every index.

If symmetry is imposed on the step biconditional, then inconsistency results. First, I will demonstrate a consequence of symmetry.

1	$A \leftrightarrow B$	Definition	1	$A \leftrightarrow B$	Definition
2	$B \leftrightarrow A$	Symmetry	2	$B \leftrightarrow A$	Symmetry
3	$\quad A^{k+1}$	hyp	3	$\quad \Box A^{k+1}$	hyp
4	$\quad B^{k}$	$\rightarrow$E	4	$\quad A^{k}$	$\Box$E
5	$\quad A^{k-1}$	$\rightarrow$E	5	$\quad B^{k+1}$	$\leftarrow$E
6	$\quad \Box A^{k}$	$\Box$I	6	$\quad A^{k+2}$	$\leftarrow$E
7	$A \rightarrow \Box A^{k+1}$	$\rightarrow$I	7	$A \leftarrow \Box A^{k+1}$	$\leftarrow$I

For the proof of inconsistency, I will use the definition $A =_{Df} \Box\sim A$.

1	$A \leftrightarrow \Box\sim A$	Definition
2	$\Box\sim A \leftrightarrow A$	Symmetry
3	$A \leftrightarrow \Box A$	Previous proof
4	$\quad A^{k+1}$	hyp
5	$\quad \Box\sim A^{k}$	$\rightarrow$E
6	$\quad \sim\Box A^{k}$	Logic of $\Box$
7	$\quad \Box A^{k}$	$\rightarrow$E
8	$\sim A^{k+1}$	RAA
9	$\sim\Box A^{k}$	$\leftarrow$Contraposition
10	A^{k+1}	$\leftarrow$E
11	$\perp^{k+1}$	$\sim$E

Imposing symmetry on the relation of definitional equivalence does not entail the corresponding material equivalence, but it can result in contradiction when certain definitions are used, namely those that "look at every other stage."

Transitivity of the step-down conditional does not hold either.

$$A \to B, B \to C \not\models A \to C$$[16]

The reason for the failure is clear when the sentences are decorated with indices. The middle term, the 'B' in the consequent of one premiss and the antecedent of the other, are not at the same level when the premises have the same index, so an argument that starts by assuming A^{i+1} will not be able to use the premises to conclude C^i, as would be needed. Instead of transitivity with the step-down conditional, a related principle is valid.

$$A \supset B, B \to C \models A \to C$$

A similar principle is valid with the step-up conditional.

$$B \leftarrow A, B \supset C \models C \leftarrow A$$

With the box, there is a valid principle that is similar to transitivity.

$$A \to B, \Box(B \to C) \models A \to \Box C$$

If definitional equivalence is transitive, then some additional definitional equivalences will entail their material equivalence counterparts. For these proofs, assume that we have a

[16]Strictly speaking, the turnstiles should have superscripts and subscripts, indicating both the set of definitions and the semantical system. These points hold for all definitions and all semantical systems I look at, so I omit the decorations.

pair of definitions $A =_{Df} B$ and $B =_{Df} C$.

1	$A \leftrightarrow B$	
2	$B \leftrightarrow C$	Definitions
3	$A \leftrightarrow C$	Transitivity
4	A^{k+2}	hyp
5	C^{k+1}	$\rightarrow$E
6	B^{k+2}	$\leftarrow$E
7	B^{k+2}	hyp
8	C^{k+1}	$\rightarrow$E
9	A^{k+2}	$\leftarrow$E
10	$A \equiv B^{k+2}$	$\equiv$I

It is difficult to see how transitivity could result in triviality, due to restrictions on the *definienda*. The *definienda* in a set of definitions are primitive predicates, rather than arbitrary formulas. In conjunction with symmetry, more definitions will produce contradictions, such as the example from the discussion of reflexivity.

Definitional equivalence, according to the revision theory, fails to be an equivalence relation. It possesses none of the three defining properties of equivalence relations. Many definitions, the ones whose revision processes settle to a fixed-point, will be equivalent to material equivalences, so those definitions will be equivalence relations. What the preceding proofs show is that if there is a distinction between definitional and material equivalence and if definitions like the ones used are to be used with classical logic, then definitional equivalence cannot be an equivalence relation.

Here are some of the principles that are valid with the step conditionals.[17]

[17]These were pointed out to me by Anil Gupta.

- $(A \to C) \supset (A \,\&\, B \to C)$
- $(A \to B) \,\&\, (A \to C) \supset .A \to (B \,\&\, C)$
- $A \vee B \to C \supset .A \to C$
- $(A \to C) \,\&\, (B \to C) \supset .A \vee B \to C$
- $(\sim A \to B) \,\&\, (\sim A \to \sim B) \supset .A$

The quantifiers interact with the step-down conditional in the same way they interact with the material conditional.

- $\vdash (\exists x Ax \to B) \equiv \forall x (Ax \to B)$, where x is not free in B
- $\vdash \forall x (B \to Ax) \equiv (B \to \forall x Ax)$, where x is not free in B
- $\vdash \forall x (Ax \to Bx) \supset (\forall x Ax \to \forall x Bx)$

There are principles that involve both step conditionals, such as the following contraposition principle.

$$(A \to B) \supset (\sim A \leftarrow \sim B)$$

Contraposition does not hold if the step conditionals are both the same step conditional. In a similar fashion, the irrelevance axiom, $A \to (B \to A)$, is not valid. Changing one of the step conditionals leads to a valid principle.

$$(B \to A) \leftarrow A$$

The two step conditionals have different logics. This may not be apparent from their rules, as the rules are so similar. One should expect some differences given the differing definitions of the two conditionals in terms of the material conditional and the box. One difference is that importation and exportation are valid for the step-up conditional but not the step-down conditional.[18]

- $\models ((C \leftarrow B) \leftarrow A) \equiv (C \leftarrow A \,\&\, B)$
- $\not\models (A \to (B \to C)) \supset A \,\&\, B \to C$
- $\not\models (A \,\&\, B \to C) \supset (A \to (B \to C))$

[18] These facts were pointed out by Anil Gupta.

There are other questions about which rules the step conditionals obey. They do not obey transitivity or substitution of equivalents.[19] The step conditionals do not obey *modus ponens*, but they obey the principle

$$A, A \to B \models \Box B.$$

The step-down conditional does not contract, but the step-up conditional does.

- $A \to (A \to B) \not\models A \to B$
- $(B \leftarrow A) \leftarrow A \models B \leftarrow A$

In the premiss, the 'B' is two stages down from the initial 'A', while in the conclusion the 'B' is only one stage away from its antecedent. Contraction will be available just for those sentences A for which we have $A \equiv \Box A$. In general, the equivalence, $A \equiv \Box A$, holds when A is $\mathscr{D}$-free.

There is a general claim worth highlighting. When A and B are $\mathscr{D}$-free, then $(A \to B) \equiv (A \supset B)$ and $(B \leftarrow A) \equiv (A \supset B)$. The step conditionals reduce to the classical material conditional when the antecedent and consequent are $\mathscr{D}$-free. This is a feature that some other approaches to adding a new conditional to a theory of truth also highlight. For example, Field points out that under the assumption of excluded middle for antecedent and consequent, his conditional is equivalent, in the sense of his new conditional, to the classical material conditional.[20] Since I am focusing on the revision theory with the classical scheme, excluded middle is not going to distinguish any formulas because the law of excluded middle is valid.

3.3 VALIDITY AND RELATED CONCEPTS

In §1.2, I defined some important concepts for revision theory, including cofinal hypotheses for a revision sequence and recurring hypotheses. I repeat the definitions here, modified to use the new revision operator for the revision theory with the box, Δ. A hypothesis h is cofinal for a revision sequence for $\Delta_{\mathscr{D},M}$ iff $\forall\alpha\exists\beta > \alpha(h = \mathscr{S}_\beta)$. A hypothesis h is recurring

[19] I will discuss substitution of equivalents more in §3.4.1.
[20] See Field (2008, 269).

for $\Delta_{\mathscr{D},M}$ iff h is cofinal in some revision sequence for $\Delta_{\mathscr{D},M}$. We can define validity for $S^{\#}$ as follows.

Definition 1 ($S^{\#}$ validity)**.** *Given a set of definitions $\mathscr{D}$, a sentence A is* valid in M on $\mathscr{D}$ *in $S^{\#}$, in symbols $M \models^{\mathscr{D}}_{\#} A$, iff for all recurring hypotheses h, there is a natural number n, such that for all $m \geq n$, A is true in $M + \Delta^{m}_{M,\mathscr{D}}(h)$. A sentence A is valid in $S^{\#}$, given $\mathscr{D}$ iff A is valid in M in $S^{\#}$ for all models M of the ground language, or in symbols $\models^{\mathscr{D}}_{\#} A$.*

Definition 2 (Entailment)**.** *Some formulas $A_1, \ldots, A_n$* entail *a formula B in $S^{\#}$ given a definition $\mathscr{D}$, $A_1, \ldots, A_n \models^{\mathscr{D}}_{\#} B$ iff $\models^{\mathscr{D}}_{\#} (A_1 \,\&\, \ldots \,\&\, A_n) \supset B$.*

Validity for $S^{\#}$ was shown not to be axiomatizable by Kremer (1993). The weaker semantical system S^0 is axiomatizable. The calculus C_0 is sound and complete with respect to the definition of validity for S_0. We need to define validity for S_0.

Definition 3 (S_0 validity)**.** *Given a set of definitions $\mathscr{D}$, a sentence A is* valid in M *in S_0, in symbols $M \models^{\mathscr{D}}_{0} A$ on $\mathscr{D}$ iff there is a natural number n, such that, for all hypotheses h, A is true in $M + \Delta^{n}_{M,\mathscr{D}}(h)$. A sentence A is valid in S_0, given $\mathscr{D}$ iff A is valid in M in S_0 for all models M of the ground language, or in symbols $\models^{\mathscr{D}}_{0} A$.*

In chapter 4, rather than use C_0, we will use a slightly modified system, $C_0^{\Box}$. Briefly, $C_0^{\Box}$ is C_0 with the addition of rules for the box and a restriction on the index shift rule, which I will explain in §4.2. We define $C_0^{\Box}$ deducibility in a way similar to C_0 deducibility from Gupta and Belnap (1993). I will use the notation $A_1^{k_1}, \ldots, A_n^{k_n} \vdash^{\mathscr{D}}_{0} B^{k_{n+1}}$ for an indexed formula being deducible from a set of indexed premises.[21] To further reduce clutter, I will use the convention that if indices on formulas are omitted, then all displayed formulas are assumed to have index 0.

With these definitions in place, I can present the relationship between the calculus $C_0^{\Box}$ and S_0.

Proposition 1. *$\vdash^{\mathscr{D}}_{0} A$ if and only if $\models^{\mathscr{D}}_{0} A$.*

The details of this proof can be found in chapter 4 (§4.2 and §4.3).

[21] I will often suppress the subscript on the turnstile to reduce clutter. I will suppress the superscript on the turnstile when context permits.

Since the step-down conditional is interdefinable with the box, the calculus resulting from adding the step-down conditional rules to C_0 is also complete.

Before moving to the philosophical discussion of the modified revision theory, I wish to draw attention to a point about finite definitions. In the original revision theory, a set of definitions was finite if there was a finite number of revisions after which any hypothesis is revised to a recurring hypothesis. This means that the revision process is over after finitely many stages in the sense that nothing new will be generated. In the present context, the old definition of finiteness will not work, because with the new definition of hypothesis, there are no finite definitions in the old sense. An example is the sequence of $\top, \Box\top, \Box^2\top, \ldots, \Box^n\top, \ldots$. Consider a hypothesis that makes $\Box^n\top$ false for each $n \geq 1$. After m revisions, $\Box^m\top$ will be true, and so be added to the next revision of the hypothesis. Since for every n, $\Delta^{n+1}(h)$ disagrees with h on at least one formula, $\Box^n\top$, there is no n at which $\Delta^n(h)$ and h agree on all formulas. For any definition, even finite definitions in the old sense, some hypotheses may have "strange" evaluations of boxed formulas, which will be filtered out after revision. No finite upper bound can be put on the revisions needed for this filtering. The problem can be fixed by changing the definitions involved, which I will do in §4.4. Roughly, a set of definitions $\mathscr{D}$ is finite, in the new sense, iff all its hypotheses are reflexive when restricted to formulas appearing in $\mathscr{D}$.

3.4 DISCUSSION

In this section, I will discuss four issues. First, I will discuss the Intersubstitutivity Principle, to set up the discussion of Field's theory in the final chapter (§3.4.1). Second, the discussion of the Intersubstitutivity Principle leads naturally to a question about whether the revision theory must reject certain argument forms so I discuss one such form (§3.4.2). Third, I will present an objection to the step conditionals that says that they are logically defective and then I will respond to this objection (§3.4.3). Finally, in §3.4.4, I will discuss the extent to which the step conditionals are compatible with a range of logical resources, which point was a major motivation for focusing on the revision theory.

3.4.1 Intersubstitutivity

One of the features of the fixed-point approach that many philosophers, including Field, want to maintain is that the truth predicate obeys the Intersubstitutivity Principle, which says that the inference from $C(A)$ to $C(T(\ulcorner A \urcorner))$, or conversely, is valid whenever C is an extensional context. Field says that truth must obey the Intersubstitutivity Principle in order to fulfill its function. A natural question is whether the truth predicate of the classical revision theory with the step-down conditional obeys the Intersubstitutivity Principle. The revision theory's truth predicate does not.

I will begin by showing that substitution of equivalents fails with the step biconditional. There are two versions of substitution of equivalents to consider. I will begin with

$$A \leftrightarrow B \models C(A) \leftrightarrow C(B).$$

To see that this fails, let the context be '$\rightarrow B$'. Then the right-hand side becomes

$$(A \rightarrow B) \leftrightarrow (B \rightarrow B).$$

The left-hand side of this biconditional is assumed to be true, while the right-hand side can, in general, fail to be true. B must contain a circularly defined predicate or operator for the right-hand side to fail. The second version of substitution of equivalents is

$$A \leftrightarrow B \models C(A) \equiv C(B).$$

A counterexample to this can be obtained by taking the context to be empty. Step equivalence between two sentences does not generally entail the material equivalence of those sentences, so this version of the substitution of equivalents is invalid.

The counterexamples to substitution of equivalents can be tailored to use the truth predicate in a way that shows the Intersubstitutivity Principle to be invalid in the revision theory. To adjust the first counterexample, let the hypothesis be the T-sentence for some paradoxical sentence, such as a Curry sentence, Tc, where $c =$ '$Tc \rightarrow \bot$'. The result of substituting this into the counterexample above yields the following conditional.

$$(T(\text{`}Tc\text{'}) \rightarrow Tc) \rightarrow (Tc \rightarrow Tc)$$

The antecedent of this conditional is one direction of the T-sentence for Tc. The consequent of this conditional is unstable, so the whole conditional is unstable. For a counterexample to the Intersubstitutivity Principle using the second form of substitution of equivalents, pick a liar sentence and use its T-sentence for the assumption of the second counterexample above.

Based on the preceding examples, we see that the revision theory's truth predicate does not obey the Intersubstitutivity Principle. This would be a problem if a theory of truth had to validate Intersubstitutivity Principle to be adequate. I will argue that theories do not have to validate the Intersubstitutivity Principle to be adequate.[22]

Many philosophers, such as Field, claim that the function or purpose of truth is to enable generalizations, particularly generalization in the sentential position. For example,with the truth predicate we can generalize from

$$\sim(\text{snow is white} \,\&\, \sim\text{snow is white})$$

to

$$\forall x \sim(T(x) \,\&\, T(\sim\!\dot{x})).$$

Without the truth predicate, we would have to use a device such as propositional quantification to achieve a similar effect and truth is generally less controversial than propositional quantification.[23] Let us look at a motivation for the adoption of the Intersubstitutivity Principle.[24]

Consider the sentence, "If everything Fred says is true, then he is in trouble." Suppose that Fred says A_1, A_2, A_3, and A_4 and he says nothing else. Let us use 'p' for "he is in trouble" and interpret the predicate F as "Fred says." There is an equivalence of some sort between

$$\forall x(Fx \supset Tx) \supset p,$$

and

$$T(\text{`}A_1\text{'}) \,\&\, T(\text{`}A_2\text{'}) \,\&\, T(\text{`}A_3\text{'}) \,\&\, T(\text{`}A_4\text{'}) \supset p.$$

[22]The argument of the next four paragraphs is based on work with Anil Gupta, who came up with the response to the Intersubstitutivity Principle to be discussed.

[23]Horsten (2009, 565 ff.) seems to suggest that propositional quantification must be understood in terms of object quantifiers and the truth predicate.

[24]For an example from the literature, see Field (2008, 210), on which this is based.

This, in turn, is equivalent, in some sense, to

$$A_1 \,\&\, A_2 \,\&\, A_3 \,\&\, A_4 \supset p.$$

There are equivalences between three sentences. Discussions of the Intersubstitutivity Principle take the two equivalences to be the same sense of equivalence. This example does not seem to require that the equivalences be the same, and the strategy of using the step conditionals denies that they are the same sense of equivalence.

To generalize the example, suppose that the predicate Bx is true of just the sentences $A_1, \ldots, A_n$. Then, for any T-free, extensional context C, there should be some sort of equivalence between the following three sentences.

(I) $C(\forall x(Bx \supset Tx))$

(II) $C(T(\text{`}A_1\text{'}) \,\&\, \ldots \,\&\, T(\text{`}A_n\text{'}))$

(III) $C(A_1 \,\&\, \ldots \,\&\, A_n)$

The generalization role of truth requires that there be some equivalence between (I) and (II) and between (II) and (III). This point is the basis of the following argument. An adequate theory of truth must provide a truth predicate that can fill the generalization role. Fulfilling the generalization role just requires the equivalences stated above. Only a theory of truth whose truth predicate obeys the Intersubstitutivity Principle can deliver the equivalences. Therefore, any theory of truth whose truth predicate violates the Intersubstitutivity Principle is inadequate as a theory of truth. I will show that this argument is unsound.

The proponent of the Intersubstitutivity Principle claims that the equivalence between the three sentences above requires the Intersubstitutivity Principle. Note that the move from (II) to (III) removes a uniform number of 'T's. A truth predicate that obeys the Intersubstitutivity Principle captures this, but the Intersubstitutivity Principle is not necessary for this conclusion. It suffices to use a weaker principle that says that iterations of the truth predicate can validly be removed or added, as in the transition from (II) to (III).

Many theories validate the addition of arbitrary iterations of the truth predicate to a T-free sentence. If the other sentences in the extensional context are T-free, such as the 'p' in the first example, then the truth predicate can be freely added to the sentences of the

wider context. A theory that validates the unrestricted addition of the truth predicate to T-free sentences is able to satisfy the demand for equivalences between the three sentence forms above.

The equivalence between (I) and (II) involves quantificational reasoning, so it is handled by material equivalence. The equivalence between (II) and (III) removes a uniform number of iterations of the truth predicate, which suggests that the equivalence is the same as that of the T-sentences, which is definitional equivalence as expressed by the step biconditional. The revision theory then maintains the equivalences between (I) and (II) and between (II) and (III) without the Intersubstitutivity Principle. The revision theory would then reject the material equivalence of (I) and (III). The revision theory satisfies the generalization role as outlined above while invalidating the Intersubstitutivity Principle, so the Intersubstitutivity Principle is not necessary for the function of truth.

While the revision theory with the step conditionals does not validate the full Intersubstitutivity Principle, it does validate a weaker principle, that of Uniform T-removal. Uniform T-Removal says that if $C(T(\ulcorner A_1 \urcorner), \ldots, T(\ulcorner A_n \urcorner))$ is an extensional context in which all the sentences with the truth predicate are displayed, then $C(T(\ulcorner A_1 \urcorner), \ldots, T(\ulcorner A_n \urcorner))$ is step equivalent to $C(A_1, \ldots, A_n)$. For example, if we have $T(\ulcorner A \urcorner) \,\&\, {\sim}T(\ulcorner B \urcorner)$, the Intersubstitutivity Principle allows one to validly infer to $T(\ulcorner A \urcorner) \,\&\, {\sim}B$. Uniform T-removal would only license the inference to $A \,\&\, {\sim}B$.

We can sharpen the formulation of when Uniform T-removal and the Intersubstitutivity Principle coincide.

Theorem 1. *Let C be any extensional, T-free context. Then for all sentences $A_1, \ldots, A_n$,*

$$C(T(\ulcorner A_1 \urcorner) \,\&\, \ldots \,\&\, T(\ulcorner A_n \urcorner)) \leftrightarrow C(A_1 \,\&\, \ldots \,\&\, A_n).$$

Proof. The proof is by induction on the complexity of the context C. The induction is straightforward. □

This theorem does not capture the full force of the Intersubstitutivity Principle, but it does capture an important sense in which the revision theory allows the substitution of A for $T(\ulcorner A \urcorner)$, and conversely. Since the revision theory rejects the Intersubstitutivity Principle,

there may be a cost in terms of arguments that the revision theory deems valid. I now turn to this.

3.4.2 Arguments

One criterion by which a descriptive theory of truth should be judged is how it evaluates arguments that use a truth predicate. A theory of truth should codify a good amount of intuitive reasoning with truth. One could strengthen this criterion to say that for simple arguments, the theory should have a simple explanation why the argument is valid or invalid. The complete proof theory can be used to demonstrate that an argument is valid for many simple arguments. The revision theory with the step conditionals performs well on the criterion of evaluating arguments.

The revision theory invalidates the Intersubstitutivity Principle. As indicated by theorem 1, the revision theory with the step conditionals captures a weakened form of the Intersubstitutivity Principle. There is a cost to rejecting the Intersubstitutivity Principle, in terms of arguments validated. The cost comes when we look at contexts that contain the truth predicate. For example, suppose that Buffy and Willow are talking and Xander is listening to them. Buffy says B_1 and Willow says W_1. Xander says that if everything Buffy said is true, then everything Willow says is true. If we interpret 'B' as "Buffy said" and 'W' as "Willow said," then we can formalize Xander's statement as follows.

(0) $\forall x(Bx \supset Tx) \supset \forall x(Wx \supset Tx)$

There is a sequence of possible inferential transitions from Xander's statement, (0).

(1) $T(\ulcorner B_1 \urcorner) \supset \forall x(Wx \supset Tx)$
(2) $\forall x(Bx \supset Tx) \supset T(\ulcorner W_1 \urcorner)$
(3) $T(\ulcorner B_1 \urcorner) \supset T(\ulcorner W_1 \urcorner)$
(4) $B_1 \supset T(\ulcorner W_1 \urcorner)$
(5) $T(\ulcorner B_1 \urcorner) \supset W_1$
(6) $B_1 \supset W_1$

There are other inferential transitions one can make from Xander's statement, but I will focus on these because they are sufficient for my point. The transitions from (0) to (1), (2),

or (3) are classically valid, given some background assumptions. The transition from (3) to (4) or (5) requires the Intersubstitutivity Principle.[25] The revision theory will invalidate the argument from (3) to (4) in general. Similarly, it will invalidate the argument from (4) to (5) in general. The argument that will be validated is the argument from (3) to (6), because this transition removes a uniform number of iterations of the truth predicate from subformulas.

The reasoning from Xander's statement to (3) to (5) has some intuitive pull, so it is a cost that the revision theory with the step conditionals evaluates it as invalid. The difference between approaches based on the Intersubstitutivity Principle and those based on uniform T-removal comes out most sharply when the wider context contains a pathological sentence or a quantified truth predicate. The truth predicate will not be eliminable nor will it be possible to add iterations of the truth predicate to the sentences of the context to permit the use of uniform T-removal. For example, let Xander assume a liar sentence, L. In that case, his assumption is equivalent to $T(\ulcorner L \urcorner) \,\&\, T(\ulcorner L \urcorner)$, which with the Intersubstitutivity Principle would yield $T(\ulcorner L \urcorner) \,\&\, L$, which is in turn equivalent to $T(\ulcorner L \urcorner) \,\&\, \sim T(\ulcorner L \urcorner)$. With the Intersubstitutivity Principle, Xander's assumption is simply inconsistent while it is just pathological with uniform T-removal.

The revision theory has a two-part response as to why the inference from (3) to (5) is appealing but generally invalid. For the first part, we need to look at instances of the inference form. If the wider context contains the truth predicate applied only to base language sentences or iterations of the truth predicate on base language sentences, then uniform T-removal can be used. This is because base language sentences all have valid generalized material T-sentences,

$$T^n(\ulcorner A \urcorner) \equiv A,$$

for each n. When W_1 is in the ground language, its step T-sentence is materially equivalent to its material T-sentence, so the argument would be valid. The transition from (3) to (5) is most attractive, I think, when it is known that W_1 is not pathological. In that case, the revision theory does accept the argument form. When W_1 contains the truth predicate, then the issue becomes more complex. There are cases in which material T-sentences for W_1 will

[25] If we are using the Intersubstitutivity Principle, the material conditional would need to be replaced with a different conditional. My discussion will ignore this detail

be valid, not all of which will be evident from the syntactic form of W_1. There are instances of the inference from (3) to (5) that are valid, but in general the inference, and similar ones, are invalid.

For the second part of the response, I will focus on the idea of definitional equivalence. The revision theory is a general theory of circular definitions, and, as indicated in §3.2.3, it distinguishes definitional equivalence from material equivalence in order to accommodate a wide range of circular definitions with classical logic. A common scenario in which the sort of inferences above are deployed is, I assume, one in which there is little danger of making a pathological attribution.[26] The appeal of the inference from (3) to (5) may, in part, be due to a failure to register the non-equivalence of step and material T-sentences. This failure may be due to an overexposure to non-circular definitions that mask the distinction between definitional and material equivalence. Alternatively, it may be due to a default expectation of non-paradoxicality. The error made in endorsing the general inference pattern found in the transition from (3) to (5) is a conflation of two notions of equivalence.

The revision theory does well evaluating other arguments.[27] It arguably does better on evaluations than the fixed-point theories. While the revision theory does not have a perfect record on evaluating arguments, I think that it does sufficiently well for the descriptive project.

3.4.3 Inadequacy

The third point I wish to discuss is an objection to the adequacy of the revision theory. The objection begins by noting that the T-sentences use the step biconditional while the semantical laws use the material biconditional. The objection says that since the biconditional used by the T-sentences is not the same as the one used by the semantical laws, none of the conditionals involved are adequate as conditionals for theories of truth. Further, the use of three conditionals adds unnecessary complications to the revision theory. This objection is not, I think, a sound objection to the revision theory with the step conditionals.

[26] As was made clear in Kripke (1975), pathological attributions occur often enough, so the assumption of non-pathologicality is at best a default.

[27] See, for example, Gupta (1984) and Kremer (2002) for discussion.

There are two points to the objection, that the step and material conditionals are individually inadequate and that the use of three conditionals adds unnecessary complications. I will take these points in turn.

The point that each conditional is inadequate is not a good objection. As I said earlier, the conditionals play different roles in the theory. The step conditionals reflect the connection of definitional equivalence.[28] Not all step biconditionals reflect definitional equivalence, but for every definitional equivalence, there is a corresponding step biconditional that is valid. This privileged set of step biconditionals reflect the particular semantic connection of definitional equivalence. The material conditional plays the role of a standard conditional. The T-sentences express a definitional equivalence, according to the revision theory, so they are properly expressed by the step biconditional. The semantical laws capture a different kind of equivalence between sentences, and they are adequately expressed with the material biconditional.

The material conditional fills the ordinary reasoning role, and, according to the revision theory view, that role does not involve expressing definitional equivalences, except in special cases. The step biconditional, in some instances, plays the role of capturing definitional equivalence. The step conditionals capture certain patterns of reasoning with circular concepts, reasoning from *definiens* to *definiendum* or conversely. The step conditionals can be used to capture ordinary reasoning in many cases, but not all. When the concepts involved are circular, the step conditionals retain their distinctive logical behavior. When the sentences involved are non-pathological and non-circular, the step conditionals reduce to the material conditional, so they can be used to codify ordinary reasoning in many contexts.[29] Ordinary reasoning with circular notions is, however, captured by the material conditional, possibly with the aid of the step conditionals. Both conditionals fill their roles adequately, and it is not an objection to point out that the conditionals have different roles.[30]

My response appeals to theory internal notion of the distinction between material and definitional equivalence. Because of this, it might not be satisfying to an objector who does

[28] I thank Anil Gupta for suggesting the 'reflects' terminology.

[29] In the system C_0, the step conditionals can stand in for the material conditional over the base language, but one must use the index shift rule, the soundness of which corresponds to the stability of the base language.

[30] If one thinks that the classical material conditional is an inadequate conditional, one can substitute a different conditional in its place.

not already adopt the revision theory, but my response is not *ad hoc* from the objector's point of view. Her main objection would be to the distinction itself, but I am not going to offer a fuller defense or motivation for that distinction here.

The second point, that the extra conditional is merely an unnecessary complication, is not a good objection. If one thinks that the revision theory offers an attractive account of truth, then the step conditionals are not a large addition. The step conditionals are, rather, already included in the philosophical picture. The revision theory holds that there are definitional equivalences, the different sets of circular definitions, and that these definitional equivalences are expressed with a connective distinct from the material biconditional. Accepting the revision theory requires accepting that there is a distinct form of non-material definitional equivalence appropriate to circular definitions. For proponents of the revision theory, the step conditionals are not a real addition. The conditionals are, rather, needed to express the definitional connections, which are already accepted.

While the addition of the step conditionals complicates the semantics, the addition does not, I think, add that much complexity.[31] The step conditionals have a simple logic with a complete calculus for reasoning about finite definitions. In short, the charge that the conditionals overly complicate the revision theory does not hold up.

As mentioned in §6.1, many approaches to truth add a new conditional to the underlying logic. In each of these approaches, there are at least two conditionals available, the added conditional and the material conditional of the underlying logic. These approaches all differ from the one suggested here in that their added conditionals are used for both the T-sentences and the semantical laws. These approaches all have two conditionals, but the primary difference is that, of their conditionals, only one has a distinguished role in each theory.

The objection, that the extra conditional is an unnecessary complication, is not a strong objection to the addition of the step conditionals to the revision theory. The step-down conditional is motivated by the overall philosophical picture of the revision theory, and it does not add much complexity to the logic.

[31] The sense of complexity employed here is non-technical.

3.4.4 Neutrality

The fourth point is that the step-down conditional is compatible with a range of logical resources. The definition of the step-down conditional I offered in §3.2.1 was for the classical scheme. The revision theory of truth is compatible with a range of logical schemes. The definition of the step-down conditional will need to be broadened to take account of the range of possible schemes. If it cannot be broadened, then adding the step-down conditional would trade the logic neutrality of the basic revision theory for the validity of the T-sentences. As argued in the previous chapter, logic neutrality should be maintained.

The following is a way to broaden the definition of the step-down conditional to other schemes.[32] Suppose that the possible semantic values of a scheme is the set $\mathcal{V}$ and that there is some set $\mathcal{D} \subseteq \mathcal{V}$ that is the set of designated values. The step-down conditional is then defined according to the following general scheme, where $a = Val(A^{i+1})$ and $b = Val(B^i)$.

A^{i+1}	B^i	$A \to B^{i+1}$
$a \in D$	$b \in D$	**d**, if $a = b$, **u** otherwise
$a \in D$	$b \in U$	**u**
$a \in U$	$b \in D$	**d**
$a \in U$	$b \in U$	**d**, if $a = b$, **u** otherwise

In this scheme, **d** is some designated value and **u** is some undesignated value. This scheme does not pin down a unique generalization, since in a many-valued logic, there may be different possibilities for the designated value to be taken in each case, as well as for the undesignated value. Such flexibility permits many cases to be distinguished in the table above, and that would provide a range of step conditionals. I am not sure whether there is anything to pick among the different conditionals in cases when different ones are available.

The definition of the step-down conditional is unusual in that it is sensitive to the structure of designatedness, rather than ordering on the values, if there is such. Here are the tables for the step-down conditional in two logics, strong Kleene and LP, that agree on ordering and number of values, but have differing designatedness structures.

[32] The definition of this section was suggested by Anil Gupta.

Example 1 (Strong Kleene step-down conditional)**.** *Here is the table for the step-down conditional defined for strong Kleene logic. This uses* ***t*** *as the designated value and* ***f*** *as the undesignated one.*

$\rightarrow$	**t**	**n**	**f**
t	**t**	**f**	**f**
n	**t**	**t**	**f**
f	**t**	**f**	**t**

Example 2 (*LP* step-down conditional)**.** *Here is the table for the step-down conditional defined for LP, again, using* ***t*** *as the designated value and* ***f*** *as the undesignated one.*

$\rightarrow$	**t**	**n**	**f**
t	**t**	**f**	**f**
n	**f**	**t**	**f**
f	**t**	**t**	**t**

The important feature of the generalized step-down conditional, together with the generalized step-up conditional, is that the step biconditional will be valid when the left-hand side is a *definiendum* with the right-hand side its *definiens*. This feature is illustrated in these examples by the diagonals of the tables taking designated values. In the classical case, the generalized scheme for the step-down conditional reduces to the definition provided in §3.2.1.

The fact that the step conditionals can be generalized to more schemes while maintaining the important property that the T-sentences are valid indicates that the expanded revision theory can maintain its logic neutrality.

There is another strategy to neutrality that is worth considering.[33] The discussion so far has been focused on the conditional. Focusing on the box leads to a slightly different approach to neutrality. Generalizing the box to non-classical schemes seems more straightforward than generalizing the step conditionals. The semantic value of $\Box A$ at stage $k+1$ is simply the semantic value of A at stage k. We can use this feature to recast the issue of step conditionals in other schemes. In classical logic, there is a single conditional, the material conditional,

[33] I thank James Shaw for suggesting this to me.

and the box is combined with that to obtain the step conditionals. In other logics, such as strong Kleene, there are conditionals besides the material conditional. The box can be combined with each of these to obtain different step conditionals. Some of the resulting step conditionals may yield a nice step biconditional, while others may not. Whether a given conditional can be used to define a good step biconditional can then become one aspect of evaluating conditionals for different schemes.

If the step-down conditional, or the box, is added to the standard revision theory, then the revision theory overcomes one of the problems it had in previous chapters, namely invalid T-sentences. The T-sentences expressed with the step biconditional are valid. There is a remaining issue of how the revision theory fairs with respect to the classification problem, the problem of classifying the semantic status of all the sentences of the language of the theory. I will now turn to this issue.

3.5 DETERMINATENESS

One of the major features of Field's theory of truth is his definition of a determinateness operator.[34] Field defines determinateness as

$$DA =_{Df} A \,\&\, {\sim}(A \rightarrow {\sim}A),$$

where '$\rightarrow$' is his conditional. Field's conditional is defined via a revision construction, and the D operator roughly says that A is $\mathbf{t}$ at the current stage and it was so at the previous stage. Welch (2008) suggests that the revision theory can mimic the determinateness operator of Field's theory. Horsten et al. (2012) develops Welch's idea by showing that the revision theory can define DA as $A \,\&\, T(\ulcorner A \urcorner)$. This has an effect similar to that of Field's determinateness operator.

For contexts in which the truth predicate is available, the definition of determinateness in terms of truth can be used. The general revision theory does not always use the truth predicate, so there are contexts in which the truth predicate is not available. In those

[34] I focus on Field's theory in the final chapter.

contexts, the revision theory can use the box to simulate the determinateness operator as $DA =_{Df} A \,\&\, \Box A$. Since the box is an operator and not a predicate, syntactic tricks involving quantification to obtain transfinite iterations generally will not be available.

Finite definitions are definitions for which the revision process is over in finitely many stages. What this means is that after finitely many revisions, the revision sequence enters a loop of finitely many hypotheses that repeat over successor stages. For these definitions, the limitation to finite iterations of the box is not an actual limitation. Further iterations would be redundant.

It may appear, then, that the revision theory can solve the classification problem using the truth predicate, and that the revision theory with the box can be used to solve a more general classification problem. This is a tempting line of thought, but I do not think that it can be maintained.

The reason that the revision-theoretic notion of determinateness cannot be used to solve the classification problem, using either the truth predicate or the box, is that the definition requires the use of *constant limit rules* and the use of a single revision sequence. A constant limit rule is one that is repeated at every limit stage. One example, the so-called Herzberger limit rule, is the rule that sets all unstable sentences to the value **f**. The Herzberger limit rule has the property that revision sequences eventually enter a loop of repeating hypotheses. This turns out to be important for the definition of determinateness.

The other feature of the definition of determinateness that I highlighted was the use of a single revision sequence. The general revision theory generates many revision sequences for a given set of definitions. Any functional allotment of semantic values to defined terms is a possible starting point for revision. One interesting set of sentences is the set that comes out as stably true, or nearly stably true, no matter what the initial hypothesis. This quantifies over all revision sequences. Another set of sentences is the set that comes out as stably true in the sequence starting from some distinguished hypothesis h. A particularly distinguished initial hypothesis sets the extension of all defined predicates to the empty set. While the empty set is a natural initial hypothesis, it is not forced by other features of the set up and the choice of the empty set as the initial hypothesis is an *ad hoc* choice.

The definition of the determinateness depends on the use of a constant, or more precisely

Herzberger, limit rule as well as the use of a single revision sequence. We can see this by showing that changing either of these parameters disturbs the consequences of the definition of determinateness.[35]

I will start with altering the limit rule. The limit rule used is the simple rule that the only sentences that receive **t** at limits are those that have stabilized at **t** sometime before the limit. We can consider a simple liar sentence, $a = \ulcorner{\sim}Ta\urcorner$, as our first example.

	0	1	2	3	...	ω	$\omega+1$	$\omega+2$
Ta	**f**	**t**	**f**	**t**		**f**	**f**	**t**
${\sim}Ta$	**t**	**f**	**t**	**f**		**f**	**t**	**f**
$T(\ulcorner Ta\urcorner)$	**f**	**f**	**t**	**f**		**f**	**f**	**t**
DTa	**f**	**f**	**f**	**f**		**f**	**f**	**f**
${\sim}DTa$	**f**	**t**	**t**	**t**		**t**	**t**	**t**

In this example, we see that Ta is unstable with a period of 2, so ${\sim}DTa$ is stably **t**. If the limit rule is changed so that all unstable sentences are set to **t** at limits, then the picture is different.

	0	1	2	3	...	ω	$\omega+1$	$\omega+2$
Ta	**f**	**t**	**f**	**t**		**t**	**t**	**f**
${\sim}Ta$	**t**	**f**	**t**	**f**		**t**	**f**	**t**
$T(\ulcorner Ta\urcorner)$	**f**	**f**	**t**	**f**		**t**	**t**	**t**
DTa	**f**	**f**	**f**	**f**		**t**	**t**	**f**
${\sim}DTa$	**f**	**t**	**t**	**t**		**t**	**f**	**t**

Unlike the previous example, ${\sim}DTa$ flips to **f** briefly after a limit stage. It stabilizes at **t** again before the next limit stage. The change is sufficient to make it so that ${\sim}DTa$ is not stable. It is, rather, nearly stable. Stability, and not near stability, will be my focus here because stability is the primary notion for Field. I focus on the behavior of stable sentences here, and these lessons can be carried over to Field's setting with a bit of work. With a small change to the limit rule, we obtain a sequence in which ${\sim}DTa$ is not stably true.

Now we look at an example for which the initial hypothesis matters more. For this example, we consider an iterated truth-teller, $b = \ulcorner T(\ulcorner Tb\urcorner)\urcorner$. Here is the table with the

[35] The ideas in the following argument are heavily indebted to the discussion of artifacts in Belnap (1982b).

initial hypothesis setting all sentences with the truth predicate to false.

	0	1	2	3	...	ω	$\omega+1$	$\omega+2$
Tb	**f**	**f**	**f**	**f**		**f**	**f**	**f**
$T(\ulcorner Tb \urcorner)$	**f**	**f**	**f**	**f**		**f**	**f**	**f**
DTb	**f**	**f**	**f**	**f**		**f**	**f**	**f**
$\sim DTb$	**f**	**t**	**t**	**t**		**t**	**t**	**t**

Here is the table setting all sentences with the truth predicate to true.

	0	1	2	3	...	ω	$\omega+1$	$\omega+2$
Tb	**t**	**t**	**t**	**t**		**t**	**t**	**t**
$T(\ulcorner Tb \urcorner)$	**t**	**t**	**t**	**t**		**t**	**t**	**t**
DTb	**t**	**t**	**t**	**t**		**t**	**t**	**t**
$\sim DTb$	**f**	**f**	**f**	**f**		**f**	**f**	**f**

In this sequence, Tb is stably true and so is DTb. Next, we consider an initial hypothesis that assigns different truth values to some sentences and at limits assigns unstable sentences the value they had at the initial hypothesis.

	0	1	2	3	...	ω	$\omega+1$	$\omega+2$
Tb	**t**	**f**	**t**	**f**		**t**	**f**	**t**
$T(\ulcorner Tb \urcorner)$	**f**	**t**	**f**	**t**		**t**	**t**	**f**
DTb	**f**	**f**	**f**	**f**		**f**	**f**	**f**
$\sim DTb$	**f**	**t**	**t**	**t**		**t**	**t**	**t**

In this sequence, neither Tb nor $T(\ulcorner Tb \urcorner)$ are stable. DTb is stably false. If we quantify over revision sequences with different initial hypotheses to obtain the set of stably true sentences, then we find that none of Tb, $T(\ulcorner Tb \urcorner)$, and DTb are in that set.

This second example changed both the limit rule and the initial hypothesis. I briefly consider a third example. Instead of an iterated truth-teller, consider a disjunction of a liar and a truth-teller, such as $c = \ulcorner Tc \urcorner$.

	0	1	2	3	...	ω	$\omega+1$	$\omega+2$
Ta	**f**	**f**	**t**	**f**		**f**	**f**	**t**
$\sim Ta$	**f**	**t**	**f**	**t**		**f**	**t**	**f**
Tc	**f**	**f**	**f**	**f**		**f**	**f**	**f**
$Ta \vee Tc$	**f**	**f**	**t**	**f**		**f**	**f**	**t**
$D(Ta \vee Tc)$	**f**	**f**	**f**	**f**		**f**	**f**	**f**
$\sim D(Ta \vee Tc)$	**f**	**t**	**t**	**t**		**t**	**t**	**t**

In this example, if the initial hypothesis assigns **t** to Tc, then $Ta \vee Tc$ will be stably true, and so will $D(Ta \vee Tc)$. This example uses the simple limit rule that assigns false to all unstable sentences, but whether something is stable or determinate depends on the initial hypothesis.

What these examples show is that altering the initial hypothesis or the limit rule may change the set of stably true sentences, and in particular it may affect whether the result of applying the D operator to some sentences stabilizes. Once we start looking at multiple sequences, it is natural to quantify over them and consider the set of sentences that stabilizes in all such revision sequences. This result in many pathological sentences that are not classified by any number of iterations of the D operator.

The notion of determinateness is sensitive to the starting parameters and the limit rule. The question is whether such a notion should count as adequate for the purposes of the classification question. While the notion of determinateness is of great technical interest, it is, I think, unable to do the philosophical job for which it has been employed.

There are three primary components of a revision sequence: the initial hypothesis, the rule of revision, and the limit rule(s). The revision rule is set by the T-sentences, and that rule is not in question. The proposal on offer picks out a particular initial hypothesis and a particular limit rule, namely the Herzberger rule.[36] These choices require strong justification, and I do not think any such can be given.

[36]The terminology "Herzberger rule" is from Gupta and Belnap (1993, 168).

The Herzberger limit rule is simple, and it has consequences that are artifacts of the choice, features that are traceable primarily to a particular *ad hoc* decision made in setting up the formal model.[37] A more complex limit rule, one that changes from limit to limit, would eliminate some of the regularities that emerge due to the simple rule.

The chosen starting hypothesis is in some ways natural, but there are many other equally natural hypotheses. Whether some sentences are stable depends on the initial hypothesis. For example, the iterated truth teller, which says of itself that it is true that it is true, is stable starting from some hypotheses and unstable starting from others. In particular, if $b = \ulcorner T(\ulcorner Tb \urcorner) \urcorner$ and the hypothesis assigns $\mathbf{f}$ to Tb and $\mathbf{t}$ to $T(\ulcorner Tb \urcorner)$, both sentences will fluctuate. If they are assigned the same value, neither will fluctuate. There seems to be no justification for starting from some particular hypothesis and not considering others as well.[38]

While the definition of determinateness may be useful in some contexts, the preceding arguments show that it is not an adequate definition of determinateness in general. The definition relies on a combination of features that is incompatible with a philosophical view naturally associated with the revision theory: in order to obtain the logic of truth for a classical language, one needs to use all revision sequences based on the T-sentences for that language.[39] This quantifies over all starting hypotheses and limit rules. The definition of determinateness does not behave well in that setting, and it yields less informative and philosophically substantive results in that setting. This definition of determinateness cannot be used to solve the classification problem for revision theory.

There is a question of whether the revision theory solves the classification problem. At each successor stage of revision, every sentence is classified as either true or false. At limit stages, the unstable sentences and their negations may fail to be in the extension of the truth predicate. Some sentences do not become nearly stable, and some do. A categoricalness predicate can be added to classify these sentences according to their stability properties across whole revision sequences. At each stage of revision for categoricalness, each sentence

[37] See Belnap (1982b) for a discussion of artifacts in the context of Herzberger's theory. I will return to this theme in the final chapter.

[38] I will return to some of the issues of limit rules and initial hypotheses in the final chapter.

[39] This view is in Belnap (1982b) and seems to be endorsed in Gupta and Belnap (1993).

will be classified as categorical or not. Some sentences will turn out to be paradoxical in a further sense that will require the use of another predicate, hyper-categorical. This process can continue. Based on my arguments from the first two chapters, I do not think this process of generating new semantic categories constitutes a problem for the revision theory's claim to satisfy the classification problem.

3.6 CONCLUSION

In this chapter, I showed how to expand the revision theory with a new conditional to address one of its flaws. I developed this proposal and defended it against objections. The revision theory augmented with the step-down conditional does well on the criteria from the first two chapters.

The addition of the step conditionals or the box point towards many logical questions. I focused mainly on the role of the step conditionals in theories of truth. The revision theory is a general theory of circular definitions, and the box has interesting applications to other aspects of the general revision theory, such as finite definitions.[40] The box can also be viewed as a modal operator with its own modal logic. These issues would take us out of the realm of theories of truth, so I will not pursue them further here.

In the next chapter, I will present the technical details of the material discussed in this chapter. Following this I will investigate the modal logic of the box. I will resume discussing the philosophical issues in the sixth and final chapter, in which I will examine Field's theory of truth. There are four reasons for focusing on Field's theory. First, Field's view is a prominent one based on the fixed-point approach, so it will provide an excellent comparison to the revision theory. Second, Field adds a conditional to the underlying logic, which in his case is strong Kleene, and there are many questions to ask about the philosophical motivations for this conditional. Third, Field's approach is motivated by issues of Semantic Closure, particularly the classification and metalanguage problems. Fourth, some of the

[40] A set of definitions is finite if all revision processes for those definitions are completed in a finite number of steps. See Gupta (2006) for more in depth discussion of finite definitions. I will discuss finite definitions more in chapter 4.

other views listed in §6.1 use relevance conditionals and are based on paraconsistent logics. Many of these theories are inconsistent but non-trivial. Field's approach and the revision theory both validate the principal of *ex falso*. The shared commitment to consistency will, I think, make the comparison cleaner.[41] The conclusion for which I will argue in the final chapter is that the revision theory with the step conditionals compares favorably to Field's theory on the criteria of the first two chapters.

[41] A detailed comparison of the revision theory with one of the other views would be valuable, but much extra philosophical work would need to be done to treat the views evenly.

4.0 EXPANDING THE REVISION THEORY

In the previous chapter, I motivated a modification of the revision theory and discussed its philosophical merits.[1] In this chapter, I will present the definitions needed for the extension of the revision theory with $\Box$. I will prove some basic results about these definitions to show that they work (§4.1). I will then prove that the proof system $C_0^{\Box}$ is sound and complete with respect to S_0 validity (§4.2-§4.3). I will then show that $C_0^{\Box}$ is complete with respect to $S^{\#}$ validity, provided $\mathscr{D}$ is a finite definition, in a sense to be defined (§4.4). Part of the proof of completeness requires showing that certain finite definitions have so-called revision indices, bounds on both the number of revisions required to reach reflexive hypotheses and on the maximum length of the cycles for reflexive hypotheses. I prove this claim in §4.5. The reader may skip this chapter, since the material in it is not needed for the philosophical discussion of the final chapter.

4.1 FOUNDATIONS

I will work with languages containing constants and variables but no function symbols. Let $\mathscr{F}$ be the set of formulas of $\mathscr{L}$ containing variables and no other terms. Let $\mathscr{F}_1$ be the set of formulas containing only variables and in which each variable occurs freely at most once in a given formula. Let $\mathscr{V}_M$ be the set of assignments of values to variables relative to a model M, although I will suppress the subscript on $\mathscr{V}$ for the rest of this section.

At times I will talk about the order of occurrence of variables. Since I am using languages with a standard syntax, the order of occurrence will be that beginning from the left side of

[1] Parts of this chapter, especially §4.1, are based on joint work with Anil Gupta.

the formula proceeding to the right.

I will use the notation o_i to refer to particular occurrences of terms in formulas. Let τ be a function such that $\tau(o) = t$ iff o is an occurrence of the term t. I will use the notation $A[t_1/o_1, \ldots, t_n/o_n]$ for the substitution of the terms $t_1, \ldots, t_n$ for the occurrences $o_1, \ldots, o_n$, assuming t_i is free for o_i. I will use $A[t_1/x_1, \ldots, t_n/x_n]$ for the simultaneous substitution of t_i for x_i. I will also assume that bound variables are relettered to avoid clashes in substitutions.

4.1.1 Similarity, hypotheses, and correspondence

We will begin with the basic definitions for similarity, hypothesis, and correspondence, together with some basic lemmas about their behavior.

Informally, similarity is a relation between two pairs of formulas and assignments. Two pairs are similar when their formulas are the same up to relettering of variables and their assignments agree on the variables "in the same positions." This is an equivalence relation. Hypotheses are special subsets of $\mathscr{F} \times \mathscr{V}$, namely those subsets closed under similarity.

We will need to use hypotheses to evaluate formulas from the whole language, not just those in $\mathscr{F}$. In order to do so, we need to define some additional concepts, the first of which is correspondence. Roughly, a pair $\langle A, v\rangle$ corresponds to $\langle B, u\rangle$ if $B \in \mathscr{F}$, B can be obtained from A by replacing names with fresh variables, and u and v agree in a certain sense. There are many useful relations between similar pairs and corresponding pairs, and we prove that they hold in the remainder of this subsection.

First, we must define the notion of an alphabetic variant. We will say that A is a one-step variant of B iff A has an occurrence a subformula $\forall x C(x)$ where B has $\forall y C(y)$, where y does not occur freely in C and y is free for x in C. A is an alphabetic variant of B just in case there is a sequence, $A = D_0, \ldots, D_n = B$, such that each D_i is a one-step variant of D_{i+1}, with the possibility that $n = 0$.[2]

Definition 4 (Similarity)**.** *Let A and B be formulas and v and v' assignments to variables. Then, $\langle A, v\rangle$ is* similar *to $\langle B, v'\rangle$ iff A and B each have exactly n occurrences of free variables*

[2]This definition is based on Hughes and Cresswell (1996, 240). See Kishida (2010, 60) for a slightly different formulation.

and there is a formula $C(x_1, \ldots, x_n) \in \mathscr{F}_1$*, whose free variables are all and only* $x_1, \ldots, x_n$*, none of which occur in* A *or* B*, such that*

1. *for some variables* $y_1, \ldots, y_n$*,* A *is an alphabetic variant of* $C(y_1, \ldots, y_n)$*,*
2. *for some variables* $z_1, \ldots, z_n$*,* B *is an alphabetic variant of* $C(z_1, \ldots, z_n)$*, and*
3. *for all* $i \leq n$*,* $v(y_i) = v'(z_i)$*.*

We note the following lemma concerning similarity.

Lemma 1. *Similarity is an equivalence relation.*

Proof. Reflexivity and symmetry are immediate. For transitivity, suppose that $\langle A, v\rangle$ is similar to $\langle B, u\rangle$ and $\langle B, u\rangle$ is similar to $\langle C, w\rangle$. Now, for some n, each of A, B and C have n occurrences of free variables. Hence, there is a formula $D(z_1, \ldots, z_n) \in \mathscr{F}_1$ and sequences $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ such that A is an alphabetic variant of $D(x_1, \ldots, x_n)$, B is an alphabetic variant of $D(y_1, \ldots, y_n)$ and for all $i \leq n$, $v(x_i) = u(y_i)$. Additionally, there is a formula $E(z'_1, \ldots, z'_n) \in \mathscr{F}_1$ and sequences $y_1, \ldots, y_n$ and $x'_1, \ldots, x'_n$ such that B is an alphabetic variant of $E(y_1, \ldots, y_n)$, C is an alphabetic variant of $E(x'_1, \ldots, x'_n)$ and for all $i \leq n$, $u(y_i) = w(x'_i)$. To obtain an appropriate formula $F(z''_1, \ldots, z''_n)$, reletter all the bound variables of E so that no bound variable occurs in A or C. If z'_i is not in A or C, let z''_i be z'_i. Otherwise, let z''_i be a new variable not occurring in A, C, or otherwise in F. Let the result of these transformations be F. Then A is an alphabetic variant of $F(x_1, \ldots, x_n)$ and C is an alphabetic variant of $F(x'_1, \ldots, x'_n)$. Finally, $v(x_i) = w(x'_i)$ since $v(x_i) = u(y_i) = w(x'_i)$. □

We can now define hypotheses as sets of pairs of formulas and assignments that are closed under similarity.

Definition 5 (Hypothesis)**.** *A* hypothesis *is a subset of* $\mathscr{F} \times \mathscr{V}$ *such that for all similar pairs* $\langle A, v\rangle$ *and* $\langle B, u\rangle$*,* $\langle A, v\rangle \in h$ *iff* $\langle B, u\rangle \in h$

In many of the definitions and proofs, we use sequences. When dealing with sequences, we will use the notation $t \in S$, where S is a sequence, to mean that for some i, $t = s_i$. We will similarly use $S \cap T = \emptyset$, where S and T are sequences, to mean that for all $s \in S$, $s \notin T$. We will use $\bar{t}$ to denote a sequence $\langle t_1, \ldots, t_n\rangle$. We will use $\bar{t}, t$ to denote a sequence

$\langle t_1, \ldots, t_n, t\rangle$. Finally, we will use the overline notation to abbreviate substitutions as follows: $A[\overline{t}/\overline{x}]$ will abbreviate $A[t_1/x_1, \ldots, t_n/x_n]$, where the n is understood from context.

The next definition will play an important role in the semantics of the extended revision theory.

Definition 6 (Corresponds)**.** *A pair* $\langle A, v\rangle$ corresponds in M *to a pair* $\langle B, v'\rangle$ *iff* $B \in \mathscr{F}$*, and there are sequences* $\langle x_1, \ldots, x_n\rangle, \langle y_1, \ldots, y_m\rangle$ *and* $\langle c_1, \ldots, c_m\rangle$ *such that* B *has exactly* $x_1, \ldots, x_n, y_1, \ldots, y_m$ *free,* $\overline{x} \cap \overline{y} = \emptyset$*, the* c_i *are all distinct names, the* y_i *are all distinct variables, and*

1. $A = B(x_1, \ldots, x_n, c_1, \ldots, c_m)$,
2. *for all* $i, 1 \leq i \leq n, v'(x_i) = v(x_i)$,
3. *for all* $i, 1 \leq i \leq m, v'(y_i) = I(c_i)$.

A formula A *corresponds in* M *to a formula* B *iff there are assignments* v *and* v' *such that* $\langle A, v\rangle$ *corresponds in* M *to* $\langle B, v'\rangle$*.*

I will frequently drop "in M" when talking about correspondence when the model is clear. I will now prove some basic facts about correspondence and its interactions with similarity.

Lemma 2. *For all* $\langle A, v\rangle$*, there is* $\langle B, v'\rangle$ *such that* $\langle A, v\rangle$ *corresponds in* M *to* $\langle B, v'\rangle$*.*

Proof. Suppose $c_1, \ldots, c_m$ are all the names in A and $x_1, \ldots, x_n$ are all the free variables of A. Let $y_1, \ldots, y_m$ be distinct variables not occurring in A. Let $B = A[y_1/c_1, \ldots, y_m/c_m]$. Let v' be the same as v except that $v'(y_i) = Val_{M,v}(c_i)$. Then by construction, $\langle A, v\rangle$ corresponds in M to $\langle B, v'\rangle$. □

Lemma 3. *Suppose that* $\langle A, v\rangle$ *corresponds in* M *to* $\langle B, v'\rangle$*. Let* $x_1, \ldots, x_n$ *and* $c_1, \ldots, c_m$ *be the sequences of free variables and names in* A *and* $y_1, \ldots, y_m$ *the free variables in* B *not occurring in* A*. Let* $o_1, \ldots, o_k$ *and* $o'_1, \ldots, o'_k$ *be the occurrences of all and only the free variables and names in* A *and* B*, respectively, in the orders of occurrence. Then,* $Val_{M,v}(\tau(o_i)) = v'(\tau(o_i))$*.*

Proof. Assume the hypotheses. Let $o_i \in \overline{o}$ be arbitrary. There are two cases: either $\tau(o_i) = x_j$ or $\tau(o_i) = c_j$, for some j.

Case: $\tau(o_i) = x_j$. Then $\tau(o'_i) = x_j$, since all variables free in A are free in the same positions in B, by the definition of correspondence in M. Then $Val_{M,v}(\tau(o_i)) = v'(\tau(o_i))$.

Case: $\tau(o_i) = c_j$. Then $\tau(o'_i) = y_j$. By the definition of corresponds in M, $Val_{M,v}(c_j) = v'(y_j)$, so $Val_{M,v}(\tau(o_i)) = v'(\tau(o_i))$.

In both cases, $Val_{M,v}(\tau(o_i)) = v'(\tau(o_i))$, as desired. □

It is also worth noting that $\langle A, v\rangle$ can correspond to $\langle B, v'\rangle$, which in turn is similar to $\langle C, v''\rangle$, but $\langle A, v\rangle$ can fail to correspond to $\langle C, v''\rangle$. The reason is that correspondence requires all occurrences of the free and bound variables of the original formula be in the corresponding formula in the same positions. It will be helpful to define a notion similar to correspondence that permits more freedom with the free variables.

Definition 7 (Pseudo-corresponds)**.** *A pair* $\langle A, v\rangle$ pseudo-corresponds in M *to a pair* $\langle B, v'\rangle$ *iff* $B \in \mathscr{F}_1$*, and there are sequences* $\langle o_1, \ldots, o_n\rangle$ *and* $\langle x_1, \ldots, x_n\rangle$*, where* $\overline{o}$ *are all and only the occurrences of free variables and names in* A *and* $\overline{x}$ *are all and only the free variables in* B*, the* x_i *are all distinct, and*

1. *No* x_i *occurs free or bound in* A
2. $A = B[\tau(o_1)/x_1, \ldots, \tau(o_n)/x_n]$,
3. *for all* $i \leq n, Val_{M,v}(\tau(o_i)) = v'(x_i)$,

Correspondence does not imply pseudo-correspondence.[3] For example, Fxx corresponds to itself, but it does not correspond to Fyz. Fxx does, however, pseudo-correspond to Fyz. Similarly, Fcc corresponds to Fxx, but Fcc does not pseudo-correspond to Fxx, as $Fxx \notin \mathscr{F}_1$.

The next three propositions show that for any corresponding pair, one can find a similar pseudo-corresponding one, and vice-versa, as illustrated in diagrams 1 and 2.[4]

[3] One could alter the definition so that correspondence implies pseudo-correspondence. We have not found any applications for such a definition of pseudo-correspondence within our framework for which the above does not adequately serve. A different set of foundational concepts could motivate such a definition. For example, one could define, relative to a given ordering of variables, the canonical form of a formula, and then define hypotheses in terms of pairs using canonical forms rather than similarity. We will not develop this idea further, so we set it aside.

[4] The convention for diagrams in this section is that continuous arrows indicate assumed relations and dotted arrows indicate relations shown to exist.

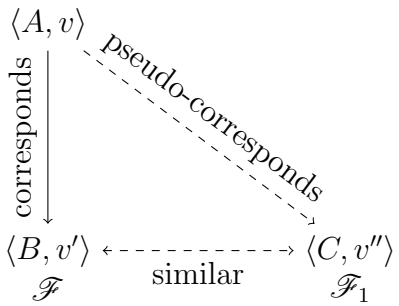

Figure 1: Corollary 1

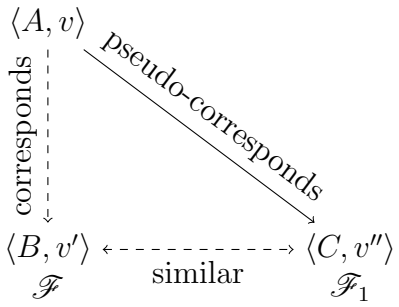

Figure 2: Lemma 5

Lemma 4. *For all $\langle A, v\rangle$ with $A \in \mathscr{F}$, there is a $B \in \mathscr{F}_1$ and $v' \in \mathscr{V}$ such that $\langle A, v\rangle$ is similar to $\langle B, v'\rangle$. Moreover, given any variables $z_1, \ldots, z_n$, such a B can be found that does not any z_i.*

Corollary 1. *Suppose $\langle A, v\rangle$ corresponds in M to $\langle B, v'\rangle$ and $B \in \mathscr{F}$. Then there is a $C \in \mathscr{F}_1$ and $v'' \in \mathscr{V}$ such that $\langle A, v\rangle$ pseudo-corresponds to $\langle C, v''\rangle$ and $\langle C, v''\rangle$ is similar to $\langle B, v'\rangle$.*

Lemma 5. *Suppose that $\langle A, v\rangle$ pseudo-corresponds to $\langle B, v'\rangle$ Then, there is some $C \in \mathscr{F}$ and $v'' \in \mathscr{V}$ such that $\langle C, v''\rangle$ is similar to $\langle B, v'\rangle$ and $\langle A, v\rangle$ corresponds in M to $\langle C, v''\rangle$.*

Proof. Assume that $\langle A, v\rangle$ pseudo-corresponds to $\langle B, v'\rangle$. By lemma 2, there is a pair $\langle C, v''\rangle$, with $C \in \mathscr{F}$, to which $\langle A, v\rangle$ corresponds in M. The similarity of $\langle B, v'\rangle$ and $\langle C, v''\rangle$ then follows by lemma 3 and noting that an appropriate formula can be found for the substitutions. □

The upshot of the preceding lemma and corollary is that instead of dealing with formulas in $\mathscr{F}$ for correspondence in M, we can deal with similar formulas in $\mathscr{F}_1$.

We note a lemma that will be used in later proofs.

Lemma 6. *If $\langle A, v\rangle$ corresponds to $\langle B, u\rangle$, then A and B have the same logical complexity.*

The following two lemmas show that correspondence is not sensitive to small details of the corresponding pair. We will say that two assignments agree on the free variables in a formula if they assign the same elements to each such variable.

Lemma 7. *If v and v' agree on all the free variables in A, then the set of pairs to which $\langle A, v\rangle$ corresponds in M is the same as the set to which $\langle A, v'\rangle$ corresponds in M.*

Lemma 8. *If $\langle A, v\rangle$ corresponds in M to both $\langle B, u\rangle$ and $\langle B', u'\rangle$, then $\langle B, u\rangle$ is similar to $\langle B', u'\rangle$.*

Lemma 9. *Suppose $\langle A, v\rangle$ and $\langle B, u\rangle$ are similar, with $\overline{c}$ all and only the names in A, then there are sequences of variables $\overline{y}$, $\overline{z}$ such that $A = A(\overline{y}, \overline{c})$ and $B = A'(\overline{z}, \overline{c})$, where A' is $A[z_1/y_1, \ldots, z_n/y_n]$, possibly with bound variables relettered.*

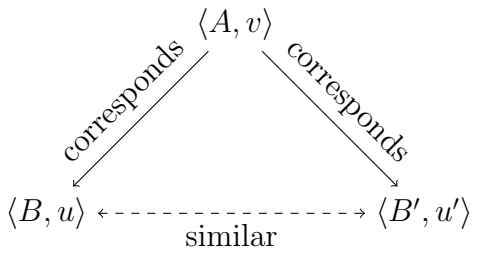

Figure 3: Lemma 8

Proof. Let $\langle A, v\rangle$ and $\langle B, u\rangle$ are similar, with $\overline{c}$ all and only the names in A. From the definition of similarity, there is a formula $C(\overline{x})$ and sequences $\overline{y}$ and $\overline{z}$ such that A is an alphabetic variant of $C(\overline{y}, \overline{c})$ and B is an alphabetic variant of $C(\overline{z}, \overline{c})$. We can reletter C so that the substitution gives $A = C(\overline{y}, \overline{c})$, which we will assume. Then we have $A = C(\overline{x}, \overline{c})[\overline{y}/\overline{x}][\overline{z}/\overline{y}] = C(\overline{y}, \overline{c})[\overline{z}/\overline{y}] = A(\overline{y}, \overline{c})[\overline{z}/\overline{y}]$, which is an alphabetic variant of $A'(\overline{z}, \overline{c}) = B$, as desired. □

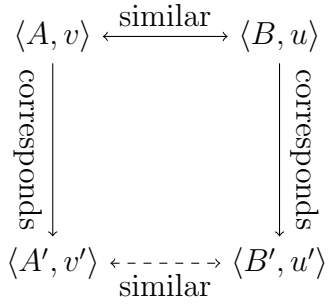

Figure 4: Lemma 10

Lemma 10. *If $\langle A, v\rangle$ and $\langle B, u\rangle$ are similar and they correspond to $\langle A', v'\rangle$ and $\langle B', u'\rangle$, respectively, then $\langle A', v'\rangle$ and $\langle B', u'\rangle$ are similar.*

Proof. Assume $\langle A, v\rangle$ and $\langle B, u\rangle$ are similar and they correspond to $\langle A', v'\rangle$ and $\langle B', u'\rangle$. From lemma 9, we may suppose that A is $A(x_1, \ldots, x_n, c_1, \ldots, c_m)$ and B is $B(y_1, \ldots, y_n, c_1, \ldots, c_m)$.

The desired conclusion then follows straightforwardly from the definitions of correspondence and similarity.

□

4.1.2 Falling under hypotheses and the semantics of box

In the previous section, we defined correspondence. In this section, we define a related concept, falling under a hypothesis, which lets us extend evaluations using hypotheses from formulas in $\mathscr{F}$ to arbitrary formulas. Roughly, a pair $\langle A, v\rangle$ falls under a hypothesis if it corresponds to some pair in the hypothesis. Once we have defined the concept of falling under a hypothesis, we can state the definition of satisfaction for formulas whose main connective is $\Box$. Following this we list some facts about correspondence and falling under hypotheses.

Definition 8 (Falling under)**.** *Let C be a formula let v be an assignment to variables. Then $\langle C, v\rangle$ falls under h relative to M, in symbols $\langle C, v\rangle \in_M h$, iff there is a pair $\langle C', v'\rangle$ such that $\langle C, v\rangle$ corresponds in M to $\langle C', v'\rangle$ and $\langle C', v'\rangle \in h$*

With the definition of a pair falling under a hypothesis relative to a model in place, I can now give the semantical clause for $\Box$ is the following.

Definition 9 (Semantics of $\Box$)**.** $M, v, h \models \Box A$ *iff* $\langle A, v\rangle \in_M h$

I will now list, without proof, some basic facts about correspondence, hypotheses, and the notion of falling under.

Lemma 11. *For $A \in \mathscr{F}$, $\langle A, v\rangle$ corresponds in M to $\langle A, v\rangle$.*

Lemma 12. *If $\langle A, v\rangle$ corresponds in M to $\langle B, u\rangle$ and $A \in \mathscr{F}$, then $A = B$.*

Lemma 13. *If $\langle A, v\rangle$ and $\langle A, u\rangle$ are similar, then $v(x) = u(x)$, for all x free in A*

Lemma 14. *For $A \in \mathscr{F}$, $\langle A, v\rangle \in_M h$ iff $\langle A, v\rangle \in h$.*

Proof. The left-to-right direction follows from lemmas 12 and 13.

The converse follows from lemma 11. □

4.1.3 Extensions

In the current framework, hypotheses are special subsets of $\mathscr{F} \times \mathscr{V}$, instead of assignments of extensions to defined predicates, as they are in the revision theory of Gupta and Belnap (1993). Extensions are useful and we want to be able to talk about the extension of a formula, in particular the extension of a defined predicate. In this subsection we define extensions and then provide the semantic clause for defined predicates. Having defined extensions, we prove a lemma showing that they are avoidable, given our other foundational definitions.

Definition 10 (Extension)**.** *Let $x_1, \ldots, x_n$ be all and only the free variables of A. Then the* extension *assigned to $A(x_1, \ldots, x_n)$ by a hypothesis h, in symbols $h(A)$, is*

$$h(A) = \{\langle d_1, \ldots, d_n\rangle \in D^n : \exists v \in V\ (v(x_1) = d_1\,\&\,\ldots\,\&\,v(x_n) = d_n\,\&\,\langle A, v\rangle \in_M h)\}.$$

When talking about extensions, I will assume that the variables under discussion are presented in the order of first free occurrence.

We now define the interpretation of defined predicates. The extension The interpretation $M + h$ is just like M except that $M + h$ assigns to the predicate G an extension based on the extension assigned to its *definiendum*. For each defined predicate $G(\overline{x}) =_{Df} A(\overline{x}, G)$ in $\mathscr{D}$,

$$Val_{M+h,v}(G) = \{\overline{d} : \exists v(v(\overline{x}) = \overline{d}\,\&\,\langle A(\overline{x}, G), v\rangle \in_M h)\}.$$

The order of the variables in the left conjunct reflects the order in which they appear in the *definiens*. The interpretation of G, $Val_{M+h,v}(G)$, will, in general, contain permutations of $h(A(\overline{x}, G))$, rather than be identical to it. If G is unary, then $Val_{M+h,v}(G)$ will be identical to $h(A(x, G))$.

We will, when dealing with the truth predicate, have to consider *partial definitions*, which are circular definitions that are defined only on particular objects.[5] If G is a unary predicate with a partial definition, let X be the set of terms on which G is defined. We will set

$$Val_{M+h,v}(G) = \bigcup_{t \in X} \{Val_{M+h,v}(t) : \langle A_t, v\rangle \in_M h\},$$

where A_t is the *definiens* for Gt.

[5] See Gupta and Belnap (1993, 197) for more on partial definitions.

I will, at times, use the notation 'M, v, h' for the interpretation $M+h$ with the assignment v. Note that on this definition of the interpretation of G, there are two "ways of interpreting G" according to a hypothesis h. One is with the extension $A(x, G)$. This is the way in which G is interpreted in a model, via $Val_{M+h,v}$. The other way, which is used primarily to evaluate $\Box G(\overline{x})$, is $h(G(\overline{x}))$. This also comes up in contexts in which we are talking about the extension assigned by h either to arbitrary formulas or to formulas from a given set.

Extensions work as one would expect. For example, $h(A(c))$ and $h(\forall y A(y))$ are both projections of $h(A(x))$, when x is free in A. The extension of $A(x, y)$ and of $A(x, y)[x/y]$ are related by the expected sort of "diagonalization" when x is free for y. Assuming no other variables are free in A, $\langle d, d\rangle \in h(A(x, y))$ just in case $d \in h(A(x, y)[x/y])$. These facts are cumbersome to state generally and to prove. We omit them here, since we will show that, in a certain sense, extensions are superfluous we can work directly with hypotheses. This streamlines later proofs.

Lemma 15 (Extension Avoidance Lemma)**.** *Let $x_1, \ldots, x_n$ be the free variables of A, in order. Then,*

$$v(\overline{x}) \in h(A) \;\Leftrightarrow\; (A, v) \in_M h.$$

Proof. For the left-to-right direction, suppose $v(\overline{x}) \in h(A)$. Then for some v', $v'(\overline{x}) = v(\overline{x})$ and $\langle A, v'\rangle \in_M h$. Since h is closed under similarity, $\langle A, v\rangle \in_M h$.

For the converse, suppose $\langle A, v\rangle \in_M h$. Therefore, there is a v' such that $v'(\overline{x}) = v(\overline{x})$ and $\langle A, v'\rangle \in_M h$, namely v. Therefore $v(\overline{x}) \in h(A)$. □

Corollary 2. *Let $G\overline{x} =_{Df} A(\overline{x}, G)$ be a definition of $\mathscr{D}$. Then,*

$$M, v, h \models G\overline{t} \;\Leftrightarrow\; \langle A(\overline{t}, G), v\rangle \in_M h.$$

4.1.4 Equality

In this subsection, we will show that hypotheses are closed under equality rules. Recall that the definition of similarity from §4.1.1 is formulated in terms of occurrences of free variables and constants. The results of this section and those of §4.1.6 are the major motivations for that definition.

The definition of similarity can be reformulated to avoid the use of occurrences, simplifying the definition and some subsequent proofs. The cost is that the following lemmas, and the semantic substitution lemma, will not hold for arbitrary hypotheses, but rather only after revision, the particular number of which depends on the modal depth of the formulas in question. In a way, such an approach is more in the spirit of the revision theory. Nonetheless, we have opted for a definition of similarity, and so of hypothesis, that preserves the soundness of all classical rules of inference with respect to arbitrary hypotheses.

Lemma 16. *Assume s is free for t in A and let A' be A with one free occurrence of s replaced by t. Suppose the following.*

1. $M, v, h \models s = t$,
2. *for all the free variables $x_1, \ldots, x_n$ common to both A and A', $v(x_i) = u(x_i)$,*
3. *if s is a variable, then $v(s) = Val_{M,u}(t)$, and*
4. *if t is a variable, then $u(t) = Val_{M,v}(s)$.*

Then, $\langle A, v\rangle \in_M h$ iff $\langle A', u\rangle \in_M h$.

Proof. Suppose s is free for t in A and $M, v, h \models s = t$. Assume that the conditions on u and v hold. Then, there is a pair, $\langle C, w\rangle$ to which $\langle A, v\rangle$ and $\langle A', u\rangle$ both pseudo-correspond. By lemma 5, $\langle A, v\rangle$ corresponds to $\langle B, v'\rangle$ and $\langle A', u\rangle$ to $\langle B', u'\rangle$, and both $\langle B, v'\rangle$ and $\langle B', u'\rangle$ are similar to $\langle C, w\rangle$. Since similarity is an equivalence relation, we conclude that $\langle A, v\rangle \in_M h$ iff $\langle A', u\rangle \in_M h$.

□

Theorem 2. *Assume s is free for t in A and let A' be A with one free occurrence of s replaced by t. Suppose the following.*

1. $M, v, h \models s = t$,
2. *for all the free variables $x_1, \ldots, x_n$ common to both A and A', $v(x_i) = u(x_i)$,*
3. *if s is a variable, then $v(s) = Val_{M,u}(t)$, and*
4. *if t is a variable, then $u(t) = Val_{M,v}(s)$.*

Then, $M, v, h \models A$ iff $M, u, h \models A'$.

Proof. Assume all the suppositions. The proof is by induction on A. The atomic cases are immediate. The classical connective cases are immediate from the induction hypothesis. We do only the box case.

Suppose A is $\Box B$. Then $M, v, h' \models \Box B$ iff $\langle B, v\rangle \in_M h$. Let B' be A' without the initial box. Then, by lemma 16, $\langle B, v\rangle \in_M h$ iff $\langle B', u\rangle \in_M h$, which is equivalent to $M, u, h \models \Box B'$, as desired. □

Define a *shifted assignment*, $v(d/x)$, to be just like v except that $v(d/x)(x) = d$.

Corollary 3. *Suppose that t is free for x in A and that $Val_{M,v}(t) = d$. Then,*

$$\langle A(x)[t/x], v\rangle \in_M h \text{ iff } \langle A(x), v(d/x)\rangle \in_M h,$$

and so

$$M, v, h \models A(x)[t/x] \text{ iff } M, v(d/x), h \models A(x).$$

Theorem 2 guarantees the soundness of the equality rule, in particular when the terms involved are within the scope of a box.

4.1.5 Revision

Next, we will define the revision operator, Δ and demonstrate that our definition is adequate. We prove that revised hypotheses are hypotheses and that revision extends to pairs containing arbitrary formulas in the expected manner. We define a useful equivalence relation on hypotheses, $\equiv_{\mathscr{D}}$, which says that two hypotheses agree on formulas appearing in the set $\mathscr{D}$. This lets us prove a kind of regularity lemma, namely that if $h \equiv_{\mathscr{D}} h'$, then $\Delta^m(h)$ and $\Delta^m(h')$ will satisfy all the same formulas of modal depth at most m. This means that, although h and h' may be wildly different with respect to most of the language, as long as they agree on formulas in $\mathscr{D}$, revision will bring them into agreement over larger fragments of the language.

The revision operator $\Delta_{M,\mathscr{D}}$ is an operation on hypotheses that satisfies the following condition for all $A \in \mathscr{F}$ and $v \in \mathscr{V}$.[6]

$$M, v, h \models A \;\Leftrightarrow\; \langle A, v\rangle \in \Delta_{M,\mathscr{D}}(h)$$

[6]We will drop subscripts on Δ when context allows.

We now demonstrate some of the relationships between hypotheses, revision, and satisfaction of formulas.

First, I will prove a lemma on assignments.

Lemma 17. *Let $A(x_1, \ldots, x_n)$ be a formula with all and only $x_1, \ldots, x_n$ free. For all assignments v, v', if, for all $i \leq n$ $v(x_i) = v'(x_i)$, then $M, v, h \models A(x_1, \ldots, x_n)$ iff $M, v', h \models A(x_1, \ldots, x_n)$.*

Proof. The proof is by induction on complexity of A. The argument is straightforward, so I will do only the box case.

Case: A is $\Box B(x_1, \ldots, x_n)$.

$$\begin{aligned} M, v, h \models \Box B(x_1, \ldots, x_n) \quad &\text{iff} \quad \langle B(x_1, \ldots, x_n), v\rangle \in_M h, \text{ by definition} \\ &\text{iff} \quad \langle B(x_1, \ldots, x_n), v'\rangle \in_M h, \text{ by lemma 7} \\ &\text{iff} \quad M, v', h \models \Box B(x_1, \ldots, x_n) \end{aligned}$$

□

Our main task is to show that revision preserves similarity. We note that correspondence and similarity both preserve satisfaction. We will omit the proofs, as both are by simple inductions.

Lemma 18. *If $\langle A, v\rangle$ corresponds in M to $\langle B, u\rangle$, then $M, v, h \models A$ iff $M, u, h \models B$.*

Lemma 19. *If $\langle A, v\rangle$ and $\langle B, u\rangle$ are similar, then $M, v, h \models A$ iff $M, u, h \models B$.*

Corollary 4. *For all hypotheses h, $\Delta(h)$ is a hypothesis.*

Lemma 19 has the following corollary, which, although obvious, is suggestive and useful.

Corollary 5. *Suppose h is a hypothesis. Let*

$$h' = \{\langle A, v\rangle \in \mathscr{F} \times \mathscr{V} : M, v, h \models A\}.$$

Then h' is a hypothesis.

The next lemma extends revision to all formulas of the language, not just those in $\mathscr{F}$.

Lemma 20. *$\langle A, v\rangle \in_M \Delta(h)$ iff $M, v, h \models A$.*

Proof. Assume $\langle A, v\rangle \in_M \Delta(h)$. So, there is a pair $\langle B, u\rangle \in \Delta(h)$ to which $\langle A, v\rangle$ corresponds. By the definition of revision, $\langle B, u\rangle \in \Delta(h)$ iff $M, u, h \models B$. By lemma 18, this is true iff $M, v, h \models A$.

Assume $M, v, h \models A$. Suppose $\langle A, v\rangle$ corresponds to $\langle B, u\rangle$. By lemma 18, $M, u, h \models B$, so $\langle B, u\rangle \in \Delta(h)$. Therefore $\langle A, v\rangle \in_M \Delta(h)$.

□

Lemma 21. *If $\langle A, v\rangle$ and $\langle B, v'\rangle$ are similar, then $\langle A, v\rangle \in_M \Delta(h)$ iff $\langle B, v'\rangle \in_M \Delta(h)$.*

Proof. This follows from lemmas 19 and 20 together with the definition of revision. □

The following is a useful definition for the upcoming proofs.

Definition 11 ($sub(\mathscr{D})$, $\equiv_{\mathscr{D}}$)**.** *Let $sub(\mathscr{D})$ be the set of subformulas of the* definienda *and the* definientia *in $\mathscr{D}$. If $B \in sub(\mathscr{D})$, we will say that B is a subformula of $\mathscr{D}$.*

Let h and h' be hypotheses. Define $h \equiv_{\mathscr{D}} h'$ iff $\forall B \in sub(\mathscr{D})$, $h(B) = h'(B)$. If $h \equiv_{\mathscr{D}} h'$, then we will say h and h' agree on $\mathscr{D}$.

The relation $\equiv_{\mathscr{D}}$ is an equivalence relation in virtue of identity being one.

Lemma 22. *For all $B \in \mathscr{F}$ and all hypotheses h and h', the following are equivalent.*

1. *for all v, $\langle B, v\rangle \in h \Leftrightarrow \langle B, v\rangle \in h'$*
2. *$h(B) = h'(B)$*

Lemma 23. *If $h \equiv_{\mathscr{D}} h'$, then for all subformulas A of $\mathscr{D}$,*

$$M, v, h \models A \;\Leftrightarrow\; M, v, h' \models A.$$

Proof. The proof is by a straightforward induction on the complexity of A.

□

Lemma 24. *If $h \equiv_{\mathscr{D}} h'$ and A is a subformula of $\mathscr{D}$, then $\Delta(h)(A) = \Delta(h')(A)$.*

Proof. The proof is by induction on the complexity of A. We will present only the box case.

Case: A is $\Box B$

$$\begin{aligned}
\langle \Box B, v\rangle \in_M \Delta(h) \quad &\text{iff} \quad M, v, h \models \Box B \\
&\text{iff} \quad \langle B, v\rangle \in_M h \\
&\text{iff} \quad \langle B, v\rangle \in_M h', \text{ by IH} \\
&\text{iff} \quad M, v, h' \models \Box B \\
&\text{iff} \quad \langle \Box B, v\rangle \in_M \Delta(h')
\end{aligned}$$

□

Corollary 6. *If $h \equiv_{\mathscr{D}} h'$, then $\Delta(h) \equiv_{\mathscr{D}} \Delta(h')$.*

Definition 12 ($\equiv_n$)**.** *Let h and h' be hypotheses. $h \equiv_n h'$ iff and for all $v \in V$ and for all B such that $d(B) \leq n$,*

$$M, v, h \models B \;\Leftrightarrow\; M, v, h' \models B$$

The $\equiv_n$ relations are equivalence relations.

Lemma 25. *For $k \leq n$, if $h \equiv_n h'$, then $h \equiv_k h'$.*

Lemma 26. *For $n \geq 0$, if $h \equiv_{\mathscr{D}} h'$ and $h \equiv_n h'$, then $\Delta(h) \equiv_{n+1} \Delta(h')$.*

Proof. Assume $h \equiv_n h'$ and $h \equiv_{\mathscr{D}} h'$. We want to show that for all formulas A such that $d(A) \leq n+1$, $\Delta(h)(A) = \Delta(h')(A)$. We proceed by induction on the complexity of formulas A. As with the previous lemma, we present only the box case.

Case: $\Box B$. Since $d(\Box B) \leq n+1$, $d(B) \leq n$. $M, v, \Delta(h) \models \Box B$ iff $\langle B, v\rangle \in_M \Delta(h)$ iff $M, v, h \models B$. By the assumption that $h \equiv_n h'$, this is equivalent to $M, v, h' \models B$, which by the definition of revision is equivalent to $\langle B, v\rangle \in_M \Delta(h')$. This holds just in case $M, v, \Delta(h') \models \Box B$, as desired.

□

Lemma 27. *For all n, if $h \equiv_{\mathscr{D}} h'$ and $h \equiv_0 h'$, then $\Delta^n(h) \equiv_n \Delta^n(h')$.*

Proof. The proof is by induction on n. The base case and induction step, respectively, are taken care of by the two preceding lemmas. □

Lemma 28. *If $h \equiv_{\mathscr{D}} h'$, then $h \equiv_0 h'$.*

Proof. Assume $h \equiv_{\mathscr{D}} h'$. We will show that for all A such that $d(A) = 0$, $M, v, h \models A$ iff $M, v, h' \models A$. The proof is by induction on the complexity of A. The cases are all trivial except for when A is a defined predicate, in which case the case is taken care of by the assumption that $h \equiv_{\mathscr{D}} h'$.

□

We can now prove an important theorem.

Theorem 3. *Suppose $h \equiv_{\mathscr{D}} h'$. If $d(A) \leq n$, then for all $m \geq n$,*

$$M, v, \Delta^m(h) \models A \textit{ iff } M, v, \Delta^m(h') \models A.$$

Proof. This follows from the two preceding lemmas. □

Corollary 7. *For all formulas A containing no* definienda*, for all n, if $d(A) \leq n$ then for all $m \geq n$,*

$$M, v, \Delta^m(h) \models A \textit{ iff } M, v, \Delta^m(h') \models A.$$

4.1.6 Semantic substitution

In this subsection, we will prove the semantic substitution lemma. We need this lemma to ensure that the quantifier rules are sound.[7] As noted in §4.1.4, this lemma is one of the major motivations for the cumbersome definition of similarity.

Lemma 29 (Semantic substitution)**.** *Suppose that t is free for x in A and that $Val_{M,v}(t) = d$. Then, $M, v, h \models A[t/x]$ iff $M, v(d/x), h \models A$.*

Proof. The proof is by induction on the complexity of A. The cases are similar to those of the proof for classical logic. We will present only the box case. Case: A is $\Box B$.

Suppose that $M, v, h \models \Box B(x)[t/x]$. This is true just in case $\langle B(x)[t/x], v\rangle \in_M h$. By corollary 3, this is equivalent to $\langle B(x), v(d/x)\rangle \in_M h$, which is in turn equivalent to $M, v(d/x), h \models \Box B(x)$.

We conclude by induction that $M, v, h \models A[t/x]$ iff $M, v(d/x), h \models A$.

□

[7] For this point, see Kishida (2010, 49-63) and Belnap and Müller (2013).

4.1.7 Definitions for revision theory

Before proceeding the soundness and completeness proofs, I will repeat some general, revision-theoretic definitions from chapter 1 and the definitions of validity from chapter 3.

Definition 13 (Revision sequence, stably in, coherence)**.** *Let On be the class of all ordinals.*

Let λ be a limit ordinal and $A \in \mathscr{F}$. $\langle A, v\rangle$ is stably in [stably out of] *$\mathscr{S}$ at λ iff*

$$\exists\alpha < \lambda \forall\beta \in [\alpha, \lambda)\langle A, v\rangle \in [\notin]\mathscr{S}_\beta$$

A hypothesis h coheres *with $\mathscr{S}$ at λ iff*

1. *if $\langle A, v\rangle$ is stably in $\mathscr{S}$ at λ, then $\langle A, v\rangle \in \mathscr{S}_\lambda$, and*
2. *if $\langle A, v\rangle$ is stably out of $\mathscr{S}$ at λ, then $\langle A, v\rangle \notin \mathscr{S}_\lambda$.*

$\mathscr{S}$ is a revision sequence *for $\mathscr{D}$ in M iff $\mathscr{S}$ is an On-long sequence of hypotheses and for all ordinals α and β,*

1. *if $\alpha = \beta + 1$, then $\mathscr{S}_\alpha = \Delta_{\mathscr{D},M}(\mathscr{S}_\beta)$, and*
2. *if α is a limit ordinal, then $\mathscr{S}_\alpha$ coheres with $\mathscr{S}$ at α.*

I will repeat some definitions for revision theory from chapter 1.

Definition 14 (Cofinal hypothesis, recurring hypothesis)**.** *A hypothesis h is* cofinal *in a revision sequence $\mathscr{S}$ for $\Delta_{\mathscr{D},M}$ iff $\forall\alpha\exists\beta \geq \alpha(\mathscr{S}_\beta = h)$.*

A hypothesis h is recurring *for $\Delta_{\mathscr{D},M}$ iff h is cofinal in some revision sequence for $\Delta_{\mathscr{D},M}$.*

As in Gupta and Belnap (1993), all definitions have cofinal and recurring hypotheses.

Finally, I will repeat the definitions of validity that I will use.

Definition 15 (S_0 validity)**.** *Given a set of definitions $\mathscr{D}$, a sentence A is* valid in M *in S_0, in symbols $M \models_0^{\mathscr{D}} A$ on $\mathscr{D}$ iff there is a natural number n, such that, for all hypotheses h, A is true in $M + \Delta^n_{M,\mathscr{D}}(h)$. A sentence A is valid in S_0, on $\mathscr{D}$ iff A is valid in M in S_0 for all models M of the ground language, or in symbols $\models_0^{\mathscr{D}} A$.*

Definition 16 ($S^\#$ validity)**.** *Given a set of definitions $\mathscr{D}$, a sentence A is* valid in M on *$\mathscr{D}$ in $S^\#$, in symbols $M \models_\#^{\mathscr{D}} A$, iff for all recurring hypotheses h, there is a natural number n, such that for all $m \geq n$, A is true in $M + \delta^m_{M,\mathscr{D}}(h)$. A sentence A is valid in S_0, on $\mathscr{D}$ iff A is valid in M in $S^\#$ for all models M of the ground language, or in symbols $\models_\#^{\mathscr{D}} A$.*

Definition 17 (Entailment)**.** *Some sentences* $A_1, \ldots, A_n$ entail *a sentence* B *in* $S^{\#}$ *given a definition* $\mathscr{D}$*,* $A_1, \ldots, A_n \models^{\mathscr{D}}_{\#} B$ *iff* $\models^{\mathscr{D}}_{\#} (A_1 \,\&\, \ldots \,\&\, A_n) \supset B$*.*

Gupta and Belnap (1993) presents a Fitch-style natural deduction proof system, C_0, for reasoning with circular definitions.[8] The system uses indices to reflect the stages of revision. All premises for classical connective rules must be at the same index, and definition rules shift the indices. We will use a modification of C_0, $C_0^{\Box}$, which was sketched in §3.3. We define $C_0^{\Box}$ in the next section. We will prove soundness of $C_0^{\Box}$ with respect to S_0 validity, and then we will prove the completeness of $C_0^{\Box}$ with respect to S_0 validity. Following this, we will prove a restricted form of completeness with respect to $S^{\#}$ validity.

4.2 SOUNDNESS

In this section, we prove the soundness of the system $C_0^{\Box}$ with respect to S_0 validity. $C_0^{\Box}$ is the proof system C_0 from Gupta and Belnap (1993, 157-160) with the addition of the following rules for the box and the index shift rule to be explained below.

$$\begin{array}{|ll} A^i & \\ \Box A^{i+1} & \Box\text{I} \end{array} \qquad \begin{array}{|ll} \Box A^{i+1} & \\ A^i & \Box\text{E} \end{array}$$

With the addition of the $\Box$ to the language, there is a question about what to do with the index shift rule. We restrict the index shift rule to formulas that contain no instances of defined expressions in $\mathscr{D}$, no instances of $\Box$, and the rule can only shift the index by one, positively or negatively, per application.[9] Clearly, an index shift rule that can change the index by arbitrary finite integers is admissible. An index shift rule for formulas that contain $\Box$ but no instances of expressions defined in $\mathscr{D}$ is admissible, so this is does not create problems. This restriction makes the soundness proof simpler. We will note the following lemma.

[8] Gupta and Belnap actually present several systems, but the one most relevant for the current project is C_0. See Gupta and Belnap (1993, 157-164) for more details and a full statement of the rules.

[9] The latter two restrictions are in place solely to make the soundness proof simpler. They can be removed at the expense of making the proof of the soundness of the index shift rule slightly more complex.

Lemma 30. *Suppose B contains no defined expressions and no instances of $\Box$. Then for all hypotheses h and h', $M, v, h \models B$ iff $M, v, h' \models B$.*

We will use the notation $\Gamma \vdash_0^{\mathscr{D}} B^k$ for the relation of there being a $C_0^{\Box}$ proof of B^k from assumptions in Γ given definitions in $\mathscr{D}$. This is defined inductively in a standard way.[10] As noted in §3.3, suppressing indices on all the sentences in a deducibility statement means that they are all have index 0. We will suppress reference to $\mathscr{D}$ in much of what follows.

We will use a more fine-grained notion of derivability than $\vdash_0^{\mathscr{D}}$. Let $\Gamma \vdash_0^{\mathscr{D},n} B^k$ be defined inductively, for each n, as the relation of there being a $C_0^{\Box}$ proof of B^k from assumptions in Γ given definitions in $\mathscr{D}$, using at most n many applications of the IS rule, DefI, DefE, $\Box$I, and $\Box$E.[11] We will state, without proof, two features of the $\vdash^{\mathscr{D},n}$ relations.

Lemma 31. *If $\Gamma \vdash^{\mathscr{D},n} B^k$, then $\Gamma \vdash^{\mathscr{D},m} B^k$, for all $m \geq n$.*

Lemma 32. *$\Gamma \vdash_0^{\mathscr{D}} B^k$ iff for some n, $\Gamma \vdash_0^{\mathscr{D},n} B^k$.*

The theoremhood in $C_0^{\Box}$ is sound with respect to S_0 validity.

Theorem 4 (Soundness)**.** *Let A be a sentence. If $\vdash_0^{\mathscr{D}} A^k$, then $\models_0^{\mathscr{D}} A$*

First, we note a small lemma.

Lemma 33. *Let $\{B_1^{m_1}, \ldots, B_j^{m_j}\}$ be a finite set of indexed sentences, let A^k be a sentence and assume $B_1^{m_1}, \ldots, B_j^{m_j} \vdash_0^{\mathscr{D}} A^k$. Then for all $n \in \mathbb{Z}$,*

$$B_1^{m_1+n}, \ldots, B_j^{m_j+n} \vdash_0^{\mathscr{D}} A^{k+n}.$$

We will now prove a stronger lemma from which soundness follows.

Lemma 34. *Let Γ be a finite set of indexed formulas, $\{A_1^{k_1}, \ldots, A_j^{k_j}\}$. Suppose that $\Gamma \vdash_0^{\mathscr{D},p} B^k$. Let m be the minimum of $k_1, \ldots, k_j, k$, and let $n_i = |k_i - m| + p$ and $n = |k - m| + p$. Then for all models M, all hypotheses h, if $M, \Delta^{n_i}(h) \models A_i$, for each $i \leq j$, then $M, \Delta^n(h) \models B$.*

Proof. Suppose $\Gamma \vdash_0^{\mathscr{D},p} B^k$. We may assume that all indices are positive.

The proof is by induction on the $\Gamma \vdash^p B^k$. The induction is straightforward. I will present a few of the cases. The proof is similar for each p. In the case in which $p = 0$, some of the

[10] See Belnap (2009, 120-121) for an example.

[11] We will frequently drop the subscript, which should always be '0'.

steps may be omitted, so we will assume $p > 0$. We will drop the superscripts to reduce clutter.

Case: $\Box$I. We are concerned with the rule leading from $\Gamma \vdash B^k$ to $\Gamma \vdash \Box B^{k+1}$. Assume that $\forall h$, if $M, v, \Delta^{n_i}(h) \models A_i$, for each $A_i \in \Gamma$, then $M, v, \Delta^n(h) \models B$. Let h be arbitrary. Assume $M, v, \Delta^{n_i}(h) \models A_i$, for each $A_i \in \Gamma$, then $M, v, \Delta^n(h) \models B$. So by the definition of revision, $\langle B, v\rangle \in_M \Delta^{n+1}(h)$, so $M, v, \Delta^{n+1}(h) \models \Box B$, by the definition of $\Box$.

Case: $\Box$E. This is similar to the $\Box$I case.

Case: DefI. We are concerned with the rule leading from $\Gamma \vdash A_G(\bar{t})^k$ to $\Gamma \vdash G(\bar{t})^{k+1}$. Assume that $\forall h$ if $M, v, \Delta^{n_i}(h) \models A_i$, for each $A_i \in \Gamma$, then $M, v, \Delta^n(h) \models A_G(\bar{t})$. Let h be arbitrary. Assume that $M, v, \Delta^{n_i}(h) \models A_i$, for each $A_i \in \Gamma$, so that $M, v, \Delta^n(h) \models A_G(\bar{t})$. By the definition of revision, $\langle A_G(\bar{t}), v\rangle \in_M \Delta^{n+1}(h)$, so $Val(\bar{t}) \in Val_{M,\Delta^{n+1}(h),v}(G)$, so $M, v, \Delta^{n+1}(h) \models G(\bar{t})$, as desired.

Case: DefE. This is similar to the DefI case.

Case: IS. This case is taken care of by lemma 30.

Case: =E. This case is taken care of by repeated applications of theorem 2.

The rest of the cases are similar to the classical case with the addition of revisions indicated in the steps so far.

□

The lemma holds for each p. If $\vdash_0^{\mathscr{D}} A^k$, then for some p, $\vdash_0^{\mathscr{D},p} A^k$. We then apply the preceding lemma to get that there is some n such that for all models M, for all hypotheses h, $M, v, \Delta^n(h) \models_0^{\mathscr{D}} A$, which yields the desired conclusion, $\models_0^{\mathscr{D}} A$.

4.3 COMPLETENESS

Next, we will prove a completeness theorem.

Theorem 5 (Completeness)**.** *If* $\models_0^{\mathscr{D}} A$, *then* $\vdash_0^{\mathscr{D}} A^0$.

Our completeness proof for $C_0^{\Box}$ of §4.2 together with a set of definitions $\mathscr{D}$ is based on the proof in Gupta and Belnap (1993), with some alterations, which I indicate below. Our proof

proceeds via a modification of the Henkin construction for the proof of the completeness of first-order logic. The primary changes are to ensure that the hypotheses are defined appropriately. We will start with some definitions.

A theory is *complete* iff it contains, for each sentence A and index i, either $A^i \in \Gamma$ or $\sim A^i \in \Gamma$. A theory is *consistent* iff $\Gamma \nvdash \bot$. A theory is *henkin* iff for all indices i and all formulas Ax with exactly one free variable, there is a closed term t such that $\exists x Ax \supset At^i \in \Gamma$.

The construction of the consistent, complete, henkin theory proceeds largely as in the standard classical case, but there are some additional closure properties to verify. Here are the closure properties of Γ with respect to $\Box$.

$$\begin{array}{lccc}
(1) & \Box A^{j+1} \in \Gamma & \text{iff} & A^j \in \Gamma \\
(2) & \Box(A \,\&\, B)^j \in \Gamma & \text{iff} & \Box A \,\&\, \Box B^j \in \Gamma \\
(3) & \Box \sim A^j \in \Gamma & \text{iff} & \sim \Box A^j \in \Gamma \\
(4) & \Box \forall x A^j \in \Gamma & \text{iff} & \forall x \Box A^j \in \Gamma \\
(5) & A^j \in \Gamma & \text{iff} & \Box^n A^j \in \Gamma, \text{ for all } n \in \mathbb{N}
\end{array}$$

In (5), A contains no expressions from $\mathscr{D}$ and no instances of $\Box$.

We note that Γ is closed under alphabetic variants and equality rules.

Lemma 35. *For all formulas A, A', for all j, if A' is an alphabetic variant of A, then $\vdash_0 \forall \overline{x}(A \equiv A')^j$.*

Proof. The proof is by a straightforward induction on the complexity of A.

□

Corollary 8. *If A and B are sentences and are alphabetic variants, then $A^j \in \Gamma$ iff $B^j \in \Gamma$.*

Lemma 36. *For all j, if $A^j \in \Gamma$ and $b = c^j \in \Gamma$, then $B^j \in \Gamma$, where B is A with 0 or more occurrences of c replaced by b.*

Next, suppose that Γ is a complete, consistent henkin theory in $C_0^{\Box}$. We will use Γ to construct a sequence of domains, the canonical model, and the canonical hypotheses,

To construct the domains, we define a relation $\equiv_j$ on the set of closed terms, $Terms$ as: $t_1 \equiv_j t_2 \;\Leftrightarrow\; (t_1 = t_2)^j \in \Gamma$. The relation $\equiv_j$ is an equivalence relation, in virtue of identity being one. Define $[a] = \{b \in Terms : a \equiv_j b\}$. The domain D_j is the set $Terms/\equiv_j$.

Since identity statements are $\mathscr{D}$- and $\Box$-free, they can be premises for the index shift rule. Therefore, $(a = b)^j \in \Gamma$ iff $(a = b)^k \in \Gamma$, for every j and k. For every j, k $D_j = D_k$.

Next, we define valuations, or assignments of values to variables. A valuation v assigns to each variable x, $v(x) \in D_j$. Since all the domains are the same, $v(x) \in D_k$ as well.

The canonical model M is constructed in stages. We begin by defining the model over the base language and then expanding it to interpret circularly defined predicates and $\Box$ with the canonical hypotheses.

Next we define the interpretation function I for the canonical model. Let j be arbitrary.

- $I(a) = [a]$, for names a.
- $I(F) = \{\overline{[t]} \in D^n : F(\bar{t})^j \in \Gamma\}$, for n-ary F.

Since constants are $\mathscr{D}$- and $\Box$-free, they will have the same interpretation in each language. Since base language atoms are $\mathscr{D}$- and $\Box$-free, index shift rules apply to them and they will have the same interpretation regardless of the j chosen.

Next we define an evaluation function $Val_{M,v}$ as follows.

- $Val_{M,v}(x) = v(x)$, for variables x.
- $Val_{M,v}(a) = I(a)$, for names a.
- $Val_{M,v}(F) = I(F)$, for atomic predicates F.

Next, we must define the satisfaction relation for the canonical model. We define satisfaction for base language atomic formulas as follows.

$$M, v \models Ft_1, \ldots, t_n \Leftrightarrow \langle Val_{M,v}(t_1), \ldots, Val_{M,v}(t_n)\rangle \in Val_{M,v}(F)$$

As $Val_{M,v}(F) = I(F)$, the right-hand side of the preceding definition is equivalent to $F(t_1, \ldots, t_n)^j \in \Gamma$. The clauses for the classical connectives and quantifiers have are standard. We will use the canonical hypotheses to interpret circularly defined predicates and boxed formulas using the revision semantics.

Next, we must the canonical hypotheses. We will use Γ to do this, but first we need some definitions.

Definition 18. *Let B be a formula whose free variables are all and only $x_1, \ldots, x_n$, $n \geq 0$. Then for $v \in V$,*

$$[B]_v = \{C \in Sent : \exists a_1 \in v(x_1) \ldots \exists a_n \in v(x_n) \,\&\, C = B[t_1/x_1, \ldots, t_n/x_n]\}$$

and

$$[B]_v^j = \{C^j : C \in [B]_v\}.$$

We will define the canonical hypotheses as follows.

Definition 19 (Canonical hypotheses)**.** $h_{j+1} = \{\langle B, v\rangle \in \mathscr{F} \times \mathscr{V} : [B]_v^j \subseteq \Gamma\}$

We will need to show that the canonical hypotheses are, in fact, hypotheses. In preparation for proving that, we will state state, without proof, some facts about the $[B]_v$s.

Lemma 37. *For all A, v, $[A]_v \neq \emptyset$.*

Lemma 38. *If $v(x) = u(x)$ for all x free in A, then $[A]_v = [A]_u$.*

Lemma 39. *If $\langle A, v\rangle$ corresponds in M to $\langle B, u\rangle$, then $[A]_v \subseteq [B]_u$.*

We will note that the converse inclusion does not hold.

Lemma 40. *For all $A \in [B]_v$, $A^j \in \Gamma$ iff $[B]_v^j \subseteq \Gamma$.*

Lemma 41. *For $A, B \in \mathscr{F}$, if $\langle A, v\rangle$ is similar to $\langle B, u\rangle$, $A' \in [A]_v$, $B' \in [B]_v$, then*

$$A'^j \in \Gamma \;\Leftrightarrow\; B'^j \in \Gamma.$$

Corollary 9. *For $A, B \in \mathscr{F}$, if $\langle A, v\rangle$ is similar to $\langle B, u\rangle$, then*

$$[A]_v^j \subseteq \Gamma \;\Leftrightarrow\; [B]_u^j \subseteq \Gamma.$$

The preceding corollary means that the canonical hypotheses are closed under similarity, and so are hypotheses. This immediately gives the following.

Corollary 10. *If $\langle A, v\rangle$ and $\langle B, v'\rangle$ are similar pairs, then $\langle A, v\rangle \in_M h_{j+1}$ iff $\langle B, v'\rangle \in_M h_{j+1}$.*

Let the model M, v, h_j agree with M, v except that h_j interprets all circularly defined predicates, G, as well as $\Box$, which receive their standard interpretations. We will prove the truth lemma with respect to the canonical model with hypotheses. For this, we will need one more lemma.

Lemma 42. *Let $A(x_1, \ldots, x_n)$ be a formula with all and only $x_1, \ldots, x_n$ free. Then, for all v, j, $M, v, h_j \models A(\overline{x})$ iff $A[a_1/x_1], \ldots, [a_n/x_n]^j \in \Gamma$, for some $a_i \in v(x_i)$.*

Proof. The proof is by induction on the construction of A. We present only some of the cases.

Base case: A is $G(\overline{t})$. Then, $M, v, h_j \models G(\overline{t})$ iff $\langle A(\overline{t}, G), v\rangle \in_M h_j$, by lemma 15. By definition, this is equivalent to $[A(\overline{t}, G)]_v^{j-1} \subseteq \Gamma$. This is the case iff $[G(\overline{t})]_v^j \subseteq \Gamma$, by the Def rules.

Case: A is $\Box B$.

$$\begin{aligned} M, v, h_{j+1} \models \Box B \quad &\text{iff} \quad \langle B, v\rangle \in_M h_{j+1} \\ &\text{iff} \quad [B]_v^j \subseteq \Gamma \\ &\text{iff} \quad [\Box B]_v^{j+1} \subseteq \Gamma, \text{ by the } \Box \text{ rules} \\ &\text{iff} \quad \Box B[\overline{a}/\overline{x}]^{j+1} \in \Gamma \end{aligned}$$

Case: A is $\forall y B$.

$$\begin{aligned} M, v, h_j \models \forall y B(\overline{x}) \quad &\text{iff} \quad \text{for all } [b],\ M, v([b]/y), h_j \models B(\overline{x}) \\ &\text{iff} \quad B[\overline{a}/\overline{x}][b/y]^j \in \Gamma, \text{ for all closed terms } b, \text{ by IH} \\ &\text{iff} \quad \forall y B[\overline{a}/\overline{x}]^j \in \Gamma \end{aligned}$$

The final equivalence is justified in two parts. From right-to-left, it is by $\forall$E. From left-to-right, it is justified by the fact that Γ is a maximal, consistent, henkin theory. Assume that $\forall y B[\overline{a}/\overline{x}]^j \notin \Gamma$. By maximality, $\exists y {\sim} B[\overline{a}/\overline{x}]^j \in \Gamma$. Since Γ is henkin, there is some constant c such that ${\sim} B[\overline{a}/\overline{x}][c/y]^j \in \Gamma$. For each closed term b, $B[\overline{a}/\overline{x}][b/y]^j \in \Gamma$, including c, which contradicts the consistency of Γ.

□

Lemma 43 (Truth lemma)**.** *For all sentences A, all assignments v, and all j, $M, v, h_j \models A \Leftrightarrow A^j \in \Gamma$*

Proof. This follows from the previous lemma. □

Next we must show that revision works properly.

Lemma 44. $\langle A, v\rangle \in_M \Delta(h_j)$ *iff* $\langle A, v\rangle \in_M h_{j+1}$.

Proof. $\langle A, v\rangle \in_M \Delta(h_j)$ if and only if $\langle B, v'\rangle \in \Delta(h_j)$, where $\langle A, v\rangle$ corresponds to $\langle B, v'\rangle$. By the definition of Δ, this is the case if and only if $M, v', h_j \models B$ if and only if $\langle B, v'\rangle \in h_{j+1}$ if and only if $\langle A, v\rangle \in_M h_{j+1}$ □

The following lemma remains.

Lemma 45. $\Delta^p(h_j) = h_{j+p}$.

Proof. This is proved by an induction on p. The base case is trivial. The previous lemma takes care of the inductive step. □

The remaining parts of the argument are carried out in much the same way as in Gupta and Belnap (1993). Suppose that A^0 is not a theorem of $C_0^{\Box}$ on the definition $\mathscr{D}$. Then $\sim A^0$ is consistent. There is a complete, consistent, henkin theory $\Gamma \supseteq \{\sim A^0\}$. Construct the canonical model for $\sim A^0$ using Γ. We must show that for each n, there is a hypothesis h such that A is false at $M + \Delta^n(h)$. For each $k \in \omega$, pick h_{-k}. By lemma 45, $\Delta^k(h_{-k}) = h_0$. By construction, A is false at $M + h_0$. We conclude that $C_0^{\Box}$ is complete with respect to validity for S_0.

4.4 FINITE DEFINITIONS

We have shown that $C_0^{\Box}$ is sound and complete with respect to S_0 validity. $S^{\#}$ validity is the more important notion, however, and it would be nice if we could extend the completeness result. Completeness is known to fail for the basic revision theory, but it holds in a special case, namely for finite definitions. The proof that the special case holds uses the notion of revision indices, which specify upper bounds on the number of revisions needed to reach recurring hypotheses and the lengths of the cycles of recurring hypotheses.[12] In §3.3, we

[12] See Gupta (2006) for details.

claimed that the old definition of finite definitions does not work in the current context. In this section, we will present new definitions for finite definitions and revision indices and then show that, if $\mathscr{D}$ is finite and has a revision index, then S_0 and $S^\#$ yield the same validities. In the next section we will prove that all finite definitions that are simple in a certain sense have revision indices.

We begin with the new definition of finite.

Definition 20 (Reflexive, finite)**.** *A hypothesis h is* n-reflexive *for $\mathscr{D}$ iff* $\forall B \in sub(\mathscr{D})\ h(B) = \Delta^n_{\mathscr{D},M}(B)$.

A hypothesis h is reflexive *for $\mathscr{D}$ iff there is some $n > 0$ for which h is n-reflexive.*

A definition $\mathscr{D}$ is finite *iff* $\forall M \exists n \forall h\ \Delta^n_{\mathscr{D},M}(h)$ *is reflexive.*

With this new definition, a set of definitions $\mathscr{D}$ is finite just in case a finite number of revisions yields hypotheses that are reflexive when restricted to $sub(\mathscr{D})$.

Reflexive hypotheses, in the original sense, are used to define a notion of finite validity, which is truth under all reflexive hypotheses. With the original definitions, if $\mathscr{D}$ is finite, then A is valid in $S^\#$ just in case it is true under every reflexive hypothesis in every model.[13] This equivalence does not hold with the new definitions. Here is a counterexample. Suppose that the set $\mathscr{D}$ contains just the definition $Gx =_{Df} Gx$. This is finite and reaches reflexive hypotheses quickly. Consider the sentence $\Box^{8000}A \vee \sim A$. Let h be a finitely reflexive hypothesis that makes all sentences with more than 8 boxes in front of them false. $\Box^{8000}A \vee \sim A$ is false in $M + h$, while it is valid in $S^\#$. The problem is that the new definition of finitely reflexive only takes into account what is going on with a small portion of the hypotheses, while testing a sentence for validity may require using a portion of the hypothesis that does not contain any formulas in $sub(\mathscr{D})$, and so is unaccounted for by the definition of finitely reflexive. This problem suggests its own solution, namely adjusting the definition to add finitely many revisions to reflexive hypotheses. This would retain the motivation behind the original definition of finite validity, that in a sense the revision process is over after finitely many revisions. The addition of extra revisions makes the notion similar to $S^\#$ validity, so we will not investigate it further here. Instead, our focus will be on the relation between S_0

[13] See Gupta (2006) for details.

validity and $S^{\#}$ validity. In the standard revision theory, all three senses of validity, $S^{\#}$, S_0, and finite, are equivalent if $\mathscr{D}$ is finite. We will show that the equivalence between S_0 and $S^{\#}$ hold under the new definitions as well.

Theorem 6. *Let $\mathscr{D}$ be a simple finite definition. Then the following are equivalent.*

- $\models_0^{\mathscr{D}} A$
- $\models_{\#}^{\mathscr{D}} A$

We will begin with the easiest part of the equivalence.

Lemma 46. *If $\models_0^{\mathscr{D}} A$, then $\models_{\#}^{\mathscr{D}}$.*

Proof. Suppose that $\mathscr{D}$ is finite and assume $\models_0^{\mathscr{D}} A$. Assume that there is some recurring hypothesis h such that for all n, there is an $m \geq n$ such that A is not true in $M + \Delta^m(h)$. Fix h. Since $\models_0^{\mathscr{D}} A$, there is a p such that for all hypotheses h', A is true in $M + \Delta^p(h')$. Fix p let $n = p$ and fix $m \geq n$. Then A is not true in $M + \Delta^m(h)$. Since $m = p + k$, let h' be $\Delta^k(h)$. Then, A is true $M + \Delta^p(\Delta^k(h))$, which is $\Delta^m(h)$. Therefore $\models_{\#}^{\mathscr{D}} A$ □

This lemma together with soundness for $C_0^{\square}$ gives the soundness theorem.

Theorem 7 (Soundness)**.** *If $\vdash_0^{\mathscr{D}} A$, then $\models_{\#}^{\mathscr{D}} A$.*

To state the next major lemma we need some definitions.

Definition 21 (Initial number, cyclic number, revision index)**.** *A natural number i is an* initial number *of $\mathscr{D}$ iff m is the least number such that for all ground models M and all hypotheses h, $\Delta^i_{M,\mathscr{D}}(h)$ is reflexive for $\mathscr{D}$.*

A natural number c is a cyclic number *of $\mathscr{D}$ iff c is the least number such that, for all ground models M and all reflexive hypotheses h, h is p-reflexive for some p such that $0 < p \leq c$.*

A revision index *for $\mathscr{D}$ is a pair $\langle i, c\rangle$ iff i is an initial number of $\mathscr{D}$ and c is a cyclic number of $\mathscr{D}$.*

The next goal is to show something close to the converse of lemma 46.

Lemma 47. *If $\mathscr{D}$ has a revision index and $\models_{\#}^{\mathscr{D}} A$, then $\models_0^{\mathscr{D}} A$.*

For the proof of this we require some lemmas.

Lemma 48. *If $\mathscr{D}$ has a revision index, then $\forall$ reflexive h $\exists$ recurring h' such that $h \equiv_{\mathscr{D}} h'$.*

Proof. Assume $\mathscr{D}$ is finite and let h be an arbitrary reflexive hypothesis for $\mathscr{D}$. So, for some k, for all subformulas B of $\mathscr{D}$, $h(B) = \Delta^k(h)(B)$. Let $\mathscr{S}$ be a revision sequence with $h = \mathscr{S}_0$ and the limit rule that unstables are set to their values in $\mathscr{S}_0$. We will show by induction that if α is a limit ordinal, then $\mathscr{S}_0 \equiv_{\mathscr{D}} \mathscr{S}_\alpha$. Assume that $\mathscr{S}_0 \equiv_{\mathscr{D}} \mathscr{S}_\beta$ for all limits $\beta < \alpha$. Let B be an arbitrary subformula of $\mathscr{D}$ and v an arbitrary valuation. There are three cases.

- $\langle B, v\rangle \in_M \mathscr{S}_\alpha$ and stably so. Then $\exists\gamma\forall\delta \in [\gamma, \alpha)\langle B, v\rangle \in \mathscr{S}_\gamma$. Fix γ, then there are two cases. If α is a successor limit, then for all successor ordinals $\delta < \alpha$, if $\delta \geq \gamma$, then $\langle B, v\rangle \in_M \mathscr{S}_\delta$. Let λ be the greatest limit $< \alpha$ or 0 if there are no such limits. By assumption, $\mathscr{S}_\lambda \equiv_{\mathscr{D}} \mathscr{S}_0$, so for some n, for all $m \geq n$, $\lambda+(m\cdot k) \geq \gamma$ and $\mathscr{S}_{\lambda+(m\cdot k)} \equiv_{\mathscr{D}} \mathscr{S}_0$. So, $\langle B, v\rangle \in_M \mathscr{S}_{\lambda+(m\cdot k)}$, so $\langle B, v\rangle \in_M \mathscr{S}_0$.

  The second case is α a limit of limits. Then there are limit ordinals β such that $\beta < \alpha$ and $\beta \geq \gamma$. Since $\langle B, v\rangle \in \mathscr{S}_\beta$ and $\mathscr{S}_\beta \equiv_{\mathscr{D}} \mathscr{S}_0$ by assumption, $\langle B, v\rangle \in_M \mathscr{S}_0$.
- $\langle B, v\rangle \notin_M \mathscr{S}_\alpha$ and stably so. This case is similar to the previous one.
- $\langle B, v\rangle$ is unstable at $\mathscr{S}_\alpha$. By the limit rule, $\langle B, v\rangle \in_M \mathscr{S}_\alpha$ iff $\langle B, v\rangle \in_M \mathscr{S}_0$

From these cases we conclude that $\langle B, v\rangle \in_M \mathscr{S}_\alpha$ iff $\langle B, v\rangle \in_M \mathscr{S}_0$. Therefore $\mathscr{S}_\alpha \equiv_{\mathscr{D}} \mathscr{S}_0$, for all limit ordinals α.

Let λ be a cofinal limit ordinal for $\mathscr{S}$. Then $\mathscr{S}_\lambda$ is recurring and $\mathscr{S}_\lambda \equiv_{\mathscr{D}} \mathscr{S}_0$, as desired. □

We can prove something close to the converse of lemma 46.

Lemma 49. *If $\mathscr{D}$ has a revision index and $\models^{\mathscr{D}}_{\#} A$, then $\models^{\mathscr{D}}_{0} A$.*

Proof. Assume that $\forall$ recurring h $\exists n\forall m \geq n$, A is true in $M + \Delta^m(h)$. Assume for reductio that $\forall p\exists h$ A is not true in $M + \Delta^p(h)$. By assumption, $\mathscr{D}$ has a revision index. Let $\langle i, c\rangle$ be a revision index of $\mathscr{D}$, where i is an initial number of $\mathscr{D}$ and c is a cyclic number of $\mathscr{D}$. Let $d(A)$ be the modal depth of A. Set $p = (i + 2) \cdot (c + 2) \cdot (d(A) + 2)$. Fix h in preparation for a reductio. Since $p \geq i$, $\Delta^p(h)$ is reflexive. By lemma 48, there is a recurring h' such that $h \equiv_{\mathscr{D}} h'$. By assumption, $\exists n\forall m \geq n$ A is true in $M + \Delta^m(h')$. Fix n. There are then two cases.

- $n \leq p$. The contradiction is immediate.
- $n > p$. Since $p, p - d(A) > i$, the initial number of $\mathscr{D}$, $\Delta^{p-d(A)}(h)$ is reflexive. There is some k such that $\Delta^{p-d(A)+k}(h) \equiv_{\mathscr{D}} \Delta^{p-d(A)}(h)$ and $p - d(A) + k > n$. We can set k to be a sufficiently large multiple of c', the cycle length for $\Delta^{p-d(A)}(h)$. By theorem 3,

$$M, v, \Delta^{p-d(A)+d(A)}(h) \models A \text{ iff } M, v, \Delta^{p-d(A)+k+d(A)}(h) \models A.$$

By assumption,

$$M, v, \Delta^{p-d(A)+d(A)}(h) \not\models A,$$

so

$$M, v, \Delta^{p-d(A)+k+d(A)}(h) \not\models A.$$

But, $p + k > n + d(A)$ and $M, v, \Delta^{p+k}(h') \models A$. Since

$$\Delta^{p+k}(h) \equiv_{\mathscr{D}} \Delta^{p+k}(h'),$$

$M, v, \Delta^{p+k}(h) \models A$, which is a contradiction.

Since both cases result in contradictions, we conclude that $\exists p \forall h\ A$ is true in $M + \Delta^p(h)$. □

Corollary 11. *If $\mathscr{D}$ has a revision index, then $\models_0^{\mathscr{D}} A$ iff $\models_{\#}^{\mathscr{D}} A$.*

In the final section, we strengthen this corollary. A *simple finite definition* is a finite definition $\mathscr{D}$ that contains finitely many *definienda*. We will show that all simple finite definitions have revision indices.

4.5 REVISION INDICES

To strengthen the preceding corollary, we now show that if $\mathscr{D}$ is a simple finite definition, then $\mathscr{D}$ has a revision index.

Theorem 8. *If $\mathscr{D}$ is a simple finite definition, then $\mathscr{D}$ has a revision index.*

Suppose that $\mathscr{D}$ consists of just $Gx =_{Df} A(x, G)$. First, we prove a useful lemma.

Lemma 50. *Suppose $\mathscr{D}_1$ is $Gx =_{Df} A(x, G)$ and $\mathscr{D}_2$ is $Gx =_{Df} A'(x, G)$, where A' is the bnf of A. Let h be a hypothesis for $\mathscr{D}_1$ and suppose $\mathscr{S}_0 = h$ is the first stage of a revision sequence over M. Then h is a hypothesis for $\mathscr{D}_2$ and for all limit ordinals λ, including 0, there is $n \in \omega$ such that for all $m \geq n$, $M, v, \mathscr{S}_{\lambda+m} \models \forall x(A \equiv A')$.*

Proof. Since the language with $\mathscr{D}_1$ is the same as that with $\mathscr{D}_2$, h is a hypothesis for $\mathscr{D}_2$. The second part follows from theorem 3. □

At limits the revision sequence can be ill-behaved, and so the equivalence may be broken for finitely many steps.

Lemma 51. *Suppose $\mathscr{D}_1$ and $\mathscr{D}_2$ are as in the preceding lemma, and h is a hypothesis for $\mathscr{D}_1$. Let $\mathscr{S}_\alpha$ be a successor stage before which $\forall x(A \equiv A')$ has stabilized to $\mathbf{t}$. Then $\Delta_{\mathscr{D}_1,M}(\mathscr{S}_\alpha) = \Delta_{\mathscr{D}_2,M}(\mathscr{S}_\alpha)$.*

Proof. Let $\mathscr{S}_\alpha$ be a successor stage before which $\forall x(A \equiv A')$ has stabilized to $\mathbf{t}$. Then $M, v, \mathscr{S}_\alpha \models \forall x(A \equiv A')$. Suppose $\langle A, v\rangle \in_M \Delta_{\mathscr{D}_1}(\mathscr{S}_\alpha)$. Then, $M, v, \mathscr{S}_\alpha \models A$, which is equivalent to $M, v, \mathscr{S}_\alpha \models A'$. This in turn is equivalent to $\langle A, v\rangle \in_M \Delta_{\mathscr{D}_2}(\mathscr{S}_\alpha)$. □

All revision sequences for $\mathscr{D}_1$ eventually become revision sequences for $\mathscr{D}_2$ over successor stages.

Corollary 12. $\models^{\mathscr{D}_1}_{\#} B$ *iff* $\models^{\mathscr{D}_2}_{\#} B$.

In these proofs, we only used a certain feature of the definitions. Let

$$d(\mathscr{D}) = sup\{n \in \omega : \exists A(A \text{ is a } \textit{definiens} \text{ in } \mathscr{D} \,\&\, d(A) = n)\}.$$

As long as $d(\mathscr{D}) < \omega$, the required equivalences are guaranteed to hold eventually. Therefore the restriction that $\mathscr{D}$ contain exactly one clause is not essential. We will assume, in light of the previous lemmas, that the boxes in the finite definitions occur only in boxed atoms.

First we need some definitions.

Definition 22. *A sequence of hypotheses $h_1, \ldots, h_n, \ldots$ for $\mathscr{D}$* does not repeat *iff for all i, j, if $i \neq j$, then $h_i \not\equiv_{\mathscr{D}} h_j$.*

A definition $\mathscr{D}$ has an ω-long [n-long] non-repeating revision sequence *iff there is a ground model M and hypothesis h such that the sequence $\langle \Delta^q_{M,\mathscr{D}}(h)\rangle_{0\leq q}$ [$\langle \Delta^q_{M,\mathscr{D}}(h)\rangle_{0\leq q\leq n}$] does not repeat.*

The main theorem we prove is the following.

Theorem 9. *If $\mathscr{D}$ is a simple finite definition, then $\mathscr{D}$ has a revision index.*

We show this by proving the following lemma.

Lemma 52. *Suppose $\mathscr{D}$ is a simple finite definition. If $\mathscr{D}$ has, for each n, a n-long non-repeating revision sequence, then $\mathscr{D}$ has an ω-long revision sequence.*

To prove this lemma, we adapt an argument from Gupta (2006). Fix a language L. Suppose that we have a finite definition $\mathscr{D}$, which just contains $Gx =_{Df} A(x, G)$. As noted, we can assume that the boxes in A occur only in boxed atoms. We can, further, assume that A contains no strings of $\Box$ applied to a non-defined predicate. As before, let $sub(\mathscr{D})$ be the set of subformulas of $\mathscr{D}$. Let j be the maximum number of boxes appearing in any boxed atom in $\mathscr{D}$. We will add to L infinitely many new unary predicates G^n_k, for $0 \leq n \leq j, k < \omega$. Each $G^n_k(x)$ will correspond to $\Delta^k(h)(\Box^n Gx)$. Call the language with the new predicates L^+ For the argument we need a few definitions.

First, define the function $f : sub(\mathscr{D}) \times \omega \to L^+$ as follows. For $k \geq 0$, define f as follows.

$$f(A,k) = \begin{cases} f(B,k)\,\&\,f(C,k) & \text{if } A = B\,\&\,C \\ \sim f(B,k) & \text{if } A = \sim B \\ \forall x f(B,k) & \text{if } A = \forall x B \\ G^n_k(t) & \text{if } A = \Box^n Gt, \\ F(\bar{t}) & \text{if } A = F(\bar{t}), \text{where } F \text{ is not a defined predicate.} \end{cases}$$

By assumption there are no formulas in $sub(\mathscr{D})$ whose major operator is $\Box$, except formulas of the form $\Box^n Gt$.

Define a sequence of sentences as follows.

E_n**:**

$$\forall x(G^0_{n+1}(x) \equiv f(A(x,G),n))$$

B^k_n**:** $0 \leq k < j$

$$\forall x(G^{k+1}_{n+1}(x) \equiv G^k_n(x))$$

C_n**:**

$$\bigwedge_{0\leq i\leq n} (\sim\forall x(G^0_{n+1} \equiv G^0_i x) \vee \ldots \vee \sim\forall x(G^j_{n+1} \equiv G^j_i x)))$$

The intuitive idea is that each E_n says that the interpretation of G^0_{n+1} is obtained using the previous stage interpretation of the *definiens*, $f(A(x,G),n)$. Each B^k_n says that the interpretation of the predicate G^k_{n+1} is obtained using the interpretation of G^{k-1}_n. Each C_n says that the interpretations of the subformulas of $\mathscr{D}$ do not repeat.

We will use a compactness argument. Define the theory Γ inductively as follows.

- $\Gamma_0 = \emptyset$
- $\Gamma_{n+1} = \Gamma_n \cup \{E_n, C_n, B^0_n, \ldots, B^j_n\}$
- $\Gamma = \bigcup_n \Gamma_n$

Next, we must show that each finite subset of Γ has a model. By compactness, Γ will have a model, which we will use to construct an ω-long non-repeating revision sequence for $\mathscr{D}$.

Assume that for each $n \geq 1$, there is a n-long, non-repeating revision sequence for $\mathscr{D}$.

Claim 52.1. *Each finite subset of* Γ *has a model.*

Proof of claim. Γ_0 is trivially consistent, so we show that each Γ_i is consistent. By assumption, for each $n \geq 1$, there is a n-long, non-repeating revision sequence for $\mathscr{D}$, so there is a model M and sequence of hypotheses $\langle \Delta^i_{\mathscr{D},M}(h)\rangle_{0\leq i\leq n}$ that does not repeat.

Let $M + \langle \Delta^i_{\mathscr{D},M}(h)\rangle_{0\leq i\leq n}$, denoted M^*, be a model that agrees with M on all the ground language symbols and interprets the new predicates as follows. For $n \geq 0$, G^n_k is interpreted as $\{v(x) \in D : M, v\ \Delta^k_{\mathscr{D},M}(h) \models \Box^n Gx\}$. These interpretations satisfy each E_n, C_n, and every B^i_n, by construction. Therefore, Γ_n is consistent. $\Box$

It follows that Γ has a model, M'. M' can be represented $M + \langle g_i \rangle_{i \in \omega}$, where M is a model of the base language L and, for each i, the g_i interpret each G_i^k. We need to transform the g_i into hypotheses. First, we explain some notation. While the model M' is $M + \langle g_i \rangle_{i \in \omega}$, it will be useful to highlight a particular g_i because it is the part that interprets a particular predicate, G_i^n. We will use the notation $M, v, g_i \models G_i^n x$ in these cases, and we will use $g_i(G_i^n)$ to pick out the set of objects satisfying G_i^n under g_i.

Let H_i be defined as follows.

- $\langle \Box^k Gx, v \rangle \in H_i$ iff $M, v, g_i \models G_i^k x$, for $k \geq 0$
- $\langle A(x, G), v \rangle \in H_i$ iff $M, v, g_i \models f(A(x, G), i)$

Define h_i as the set of pairs similar to a pair to which some $(B, v) \in H_i$ corresponds in M'. We need to show that this definition of the h_is work.

Claim 52.2. *Each h_i is closed under similarity.*

Proof of claim. Immediate from the definition. □

We turn the extensions assigned by M' to the new predicates of L^+ into a series of hypotheses for L with the defined predicate G. For each g_i, we use the previous lemma to obtain a hypothesis that is equivalent over the relevant formulas, which are all members of $sub(\mathscr{D})$. As the g_is are defined only over the indicated predicates, the h_is assign empty extensions to most formulas of L^+. Next, we must transform the sequence $\langle h_i \rangle_{i \in \omega}$ into a revision sequence for $\mathscr{D}$.

Claim 52.3. *For all $B \in sub(\mathscr{D})$, $M, v, h_n \models B$ iff $M, v, g_{n+1} \models f(B, n+1)$.*

Proof of claim. The proof is by induction on B. We present only the interesting cases.

Case: B is Gx.

$$\begin{aligned} M, v, h_n \models Gx \quad &\text{iff} \quad \langle A(x, G), v \rangle \in_M h_n, \text{ lemma 15} \\ &\text{iff} \quad M, v, g_n \models f(A, n), \text{ by definition} \\ &\text{iff} \quad M, v, g_{n+1} \models G_{n+1}^0 x, \text{ by definition of } \Gamma \end{aligned}$$

Case: B is $\Box C$. Then B is $\Box^{k+1} Gx$ for some $k \geq 0$.

$$\begin{aligned} M, v, h_n \models \Box^{k+1} Gx \quad & \text{iff} \quad \langle \Box^k Gx, v\rangle \in h_n \\ & \text{iff} \quad M, v, g_n \models G_n^k x, \text{ by definition} \\ & \text{iff} \quad M, v, g_{n+1} \models G_{n+1}^{k+1} x, \text{ by definition of } \Gamma \end{aligned}$$

□

Let $X = \{B : B \text{ has either the form } \Box^k Gx, \text{ for some } k \leq d(\mathscr{D}), \text{ or } A(x, G)\}$.

Claim 52.4. *For $B \in X$, $M, v, h_n \models B$ iff $\langle B, v\rangle \in_M h_{n+1}$.*

Proof of claim. There are two cases: either B is $\Box^k Gx$, for $k \geq 0$, or B is $A(x, G)$.

Suppose B is $\Box^k Gx$.

$$\begin{aligned} M, v, h_n \models \Box^k Gx \quad & \text{iff} \quad M, v, g_{n+1} \models G_{n+1}^k x, \text{ by claim 52.3} \\ & \text{iff} \quad \langle \Box^k Gx, v\rangle \in h_{n+1}, \text{ by definition of } \Gamma \end{aligned}$$

Suppose B is $A(x, G)$.

$$\begin{aligned} M, v, h_n \models A(x, G) \quad & \text{iff} \quad M, v, g_{n+1} \models f(A, n+1), \text{ by claim 52.3} \\ & \text{iff} \quad \langle A(x, G), v\rangle \in_M h_{n+1}, \text{ by definition} \end{aligned}$$

□

Over the formulas in X, the h_i appear close to a revision sequence. All that is left to do is embed them in a revision sequence.

Claim 52.5. *Suppose $\langle h_n\rangle_{n\in\omega}$ is a sequence of hypotheses such that for all n,*

$$\forall B \in X(M, v, h_n \models B \Leftrightarrow \langle B, v\rangle \in_M h_{n+1}).$$

Then there is a sequence of hypotheses $\langle h'_n\rangle_{n\in\omega}$ satisfying the following:

(1) $\forall A(M, v, h'_n \models A \Leftrightarrow \langle A, v\rangle \in_M h'_{n+1})$,
(2) $\forall B \in sub(\mathscr{D})(M, v, h_n \models B \Leftrightarrow M, v, h'_n \models B)$, *and*
(3) $\forall B \in X(\langle B, v\rangle \in_M h'_n \Leftrightarrow \langle B, v\rangle \in_M h_n)$,

Proof of claim. Let $\langle h_n\rangle_{n\in\omega}$ be a sequence of hypotheses satisfying the assumptions. Define a new sequence of hypotheses inductively as follows.

- $h'_0 = h_0$
- $h'_{n+1} = \{\langle A, v\rangle \in \mathscr{F} \times \mathscr{V} : M, v, h'_n \models A\}$

It remains to show that this sequence works. Clearly property (1) holds. We need to show that (2) and (3) hold and that each h'_i is a hypothesis.

By definition, h'_0 is a hypothesis. By lemma 20, each h'_n, $n \geq 1$, is a hypothesis.

We show that (2) holds with a proof by a double induction on n and the complexity of B. The base case is trivial.

Suppose (2) holds for n, and we show it for $n+1$.

Base case: B is a base language atomic sentence. This is trivial.

Base case: B is Gx.

$$\begin{array}{lll} M, v, h_{n+1} \models Gx & \text{iff} & \langle A(x, G), v\rangle \in_M h_{n+1}, \text{ by lemma 15} \\ & \text{iff} & M, v, h_n \models A(x, G), \text{ by claim 52.4} \\ & \text{iff} & M, v, h'_n \models A(x, G), \text{ by the outer induction hypothesis} \\ & \text{iff} & M, v, h'_{n+1} \models Gx \end{array}$$

The inductive cases for $C \,\&\, D$, $\sim C$, and $\forall x C$ are all handled by the inner induction hypothesis.

Case: B is $\Box C$. Then, B is $\Box^{k+1} Gx$, for some $k \geq 0$.

$$\begin{array}{lll} M, v, h_{n+1} \models \Box^{k+1} Gx & \text{iff} & \langle \Box^k Gx, v\rangle \in_M h_{n+1}, \text{ by definition} \\ & \text{iff} & M, v, h_n \models \Box^k Gx, \text{ by claim 52.4} \\ & \text{iff} & M, v, h'_n \models \Box^k Gx, \text{ by the outer induction hypothesis} \\ & \text{iff} & M, v, h'_{n+1} \models \Box^{k+1} Gx \end{array}$$

Finally, given the definition of $\langle h'_i\rangle_{i\in\omega}$, (2) implies (3), so we are done.

□

The sequence of hypotheses $\langle h'_i\rangle_{i\in\omega}$ is an ω-long non-repeating revision sequence for $\mathscr{D}$ in virtue of Γ containing C_n, for each n. This completes the proof of lemma 52.

The proof that we gave below required that $\mathscr{D}$ be a simple finite definition for the construction of Γ. We think that it can be generalized to permit $\mathscr{D}$ to contain infinitely many clauses. This raises many related questions about finite definitions with infinitely many clauses. We will not pursue those here.

We can now prove the main theorem of this section, that $\mathscr{D}$ has a revision index.

Lemma 53. *If $\mathscr{D}$ is a simple finite definition, then $\mathscr{D}$ has an initial number.*

Proof. Assume the contrary. Then, for each n, there is a model M and n-long non-repeating revision sequence. By lemma 52, there is a model M' and and ω-long non-repeating revision sequence for $\mathscr{D}$. This contradicts the assumption that $\mathscr{D}$ is finite. Therefore $\mathscr{D}$ has an initial number. □

Lemma 54. *If $\mathscr{D}$ is a simple finite definition, then $\mathscr{D}$ has an initial number.*

The proof is similar to that of the previous lemma. These two lemmas give us the desired theorem, that $\mathscr{D}$ has a revision index.

In light of the earlier completeness theorem, we have the following.

Theorem 10 (Completeness for finite definitions)**.** *Suppose $\mathscr{D}$ is a simple finite definition. If $\models^{\mathscr{D}}_{\#} A$ then $\vdash^{\mathscr{D}}_{0} A$.*

In this chapter we have presented the basic definitions needed for the revision theory expanded with the box and demonstrated some of their basic properties. We then proved soundness and completeness results for the proof system $C_0^{\Box}$ with respect to S_0 validity. This led to a question about the relation of $C_0^{\Box}$ to $S^{\#}$. We showed that S_0 is equivalent to $S^{\#}$ over finite definitions with revision indices, and then we showed that, in fact, all simple finite definitions have revision indices. This leaves open several questions, such as: Do all finite definitions have revision indices? Are non-simple finite definitions equivalent, in some sense, to a combination of simple finite definitions? We will leave these questions for future investigation. We will now examine the modal logic of the box in more detail.

5.0 THE MODAL LOGIC OF REVISION

The box was introduced in chapter 3 with a specific interpretation in the revision theory. We can, however, view the box as a modal operator and investigate its modal logic. The logic is simple, but nonetheless has many interesting features. We will present two proof systems for the logic, one for the propositional logic and one for the first-order logic (§5.1). We will prove equivalences between our different formulations and show that the systems behave nicely in other respects. We will then prove a completeness result relating the modal logic and finite definitions in the revision theory (§5.2). The reader can skip this chapter without loss of continuity.

5.1 MODAL LOGIC

We will begin with some definitions to establish terminology. We will use the standard definitions of Kripke frames and Kripke models. For the propositional case, we will use the notation, $M, w \Vdash p$, for p being true at w in the Kripke model M. The '$\Vdash$' relation is extended to complex sentences in the standard way. For the first-order case, the '$\Vdash$' notation will be adapted in the obvious way. We will use only first-order Kripke models with constant domains.

The following is a bit of terminology that will be useful.

Definition 23. *A formula A is a* boxed atom *iff for some $n \geq 0$, it has the form $\Box^n p$, where p is an atom.*

A maximal boxed atom *is an occurrence of a boxed atom that is not a subformula of*

another boxed atom.

As an example, in the sentence $\Box^4 p \,\&\, \Box^2 p \,\&\, p$ there are three maximal boxed atoms, $\Box^4 p$, $\Box^2 p$, and p, although both of the latter two appear as subformulas of the first. We will generally restrict attention to maximal boxed atoms.

The modal logic we will investigate is obtained by adding to the modal logic K the axiom

$$\Diamond A \equiv \Box A.^{1}$$

The resulting logic will be RT, for "revision theory." The logic is sound and complete for frames in which every world has exactly one R-successor. From a certain perspective, these structures look like revision sequences. There is something to this idea, but there is an important difference. In revision sequences, the box and the "accessibility relation" of the revision sequence go in the same direction. If p is true at stage k, $\Box p$ is true at stage $k+1$. In the possible worlds models, the box looks down the accessibility relation. If p is true at w, then for all u such that uRw, $\Box p$ is true at u. This difference does not undermine the noted analogy, as I will show in §5.2.

I will list some theorems of RT that will be useful later on.

Lemma 55. *RT proves the following.*

1. $\Box{\sim}A \equiv {\sim}\Box A$
2. $\Box(A \supset B) \equiv (\Box A \supset \Box B)$
3. $\Box(A \vee B) \equiv (\Box A \vee \Box B)$
4. $\Box A \vee \Box{\sim}A$
5. ${\sim}\Box\bot$
6. $\bot \equiv \Box^n \bot$, *for all* n
7. $\Box^n(A \supset B) \equiv (\Box^n A \supset \Box^n B)$, *for all* n.

Besides a Hilbert system, we can also provide a sequent system for RT. We will call this sequent system LRT. For the purposes of the sequent system, we will treat '$\Diamond$' purely as a defined symbol, with the definition '${\sim}\Box{\sim}$'. We take a standard multiple conclusion sequent calculus for classical logic and add the following rule.

[1] The $\Box$ only version is ${\sim}\Box A \equiv \Box{\sim}A$.

$$\frac{\Gamma \vdash \Delta}{\Box\Gamma \vdash \Box\Delta}\,(\Box)$$

As is common, if Γ is $A_1, \ldots, A_n$, then $\Box\Gamma$ is $\Box A_1, \ldots, \Box A_n$. This looks similar to the modal rule for a sequent system for K, but that rule requires that Γ and Δ be non-empty. Otherwise, the following is derivable.

$$\dfrac{\dfrac{\dfrac{A \vdash A}{\vdash A, \sim A}\,(\vdash\sim)}{\vdash \Box A, \Box\sim A}\,(\Box)}{\vdash \Box A \vee \Box\sim A}\,(\vdash\vee)$$

The final sequent is invalidated by a K model. Let $W = \{w, u, v\}$ and suppose wRu and wRv. Let $u \Vdash A$ and $v \Vdash \sim A$. Then $w \not\Vdash \Box A \vee \Box\sim A$. This sentence is valid in RT, however. Using the $(\Box)$ rule, we can derive the commutation of the box with all the logical connectives. As an illustration, here is one of the derivations.

$$\dfrac{\dfrac{\dfrac{\dfrac{A \vdash A, B}{\vdash A, A \supset B}\,(\vdash\supset)}{\vdash \Box A, \Box(A \supset B)}\,(\Box) \qquad \dfrac{\dfrac{B, A \vdash B}{B \vdash A \supset B}\,(\vdash\supset)}{\Box B \vdash \Box(A \supset B)}\,(\Box)}{\Box A \supset \Box B \vdash \Box(A \supset B), \Box(A \supset B)}\,(\supset\vdash)}{\Box A \supset \Box B \vdash \Box(A \supset B)}\,(\vdash W)$$

One sign that a sequent system is well-behaved is that it permits an elimination theorem.

Theorem 11 (Cut Elimination)**.** *Cut is an admissible rule for the sequent system LRT.*

Proof. The proof proceeds via the standard technique of using a double induction on degree and rank of the cut formula. We will calculate rank in the usual way except that after an application of the $(\Box)$ rule, all sentences have rank 0. As is common, instead of cut, the stronger mix rule is used. I will use the notation Γ^{-A} for the set Γ with all instances of A removed.[2] Since the system is the classical sequent calculus with the additional modal rule, I will omit the old cases. There is one new case to check, the case in which the cut formula is $\Box A$ and it was just introduced.

$$\frac{\dfrac{\Gamma \vdash \Delta, A}{\Box\Gamma \vdash \Box\Delta, \Box A}\,(\Box) \qquad \dfrac{A, \Sigma \vdash \Theta}{\Box A, \Box\Sigma \vdash \Box\Theta}\,(\Box)}{\Box\Gamma, (\Box\Sigma)^{-\Box A} \vdash (\Box\Delta)^{-\Box A}, \Box\Theta}\,(\text{mix})$$

The mix can be carried out on simpler formulas as follows.

[2]This notation is due to Nuel Belnap.

$$\dfrac{\dfrac{\Gamma \vdash \Delta, A \qquad A, \Sigma \vdash \Theta}{\Gamma, \Sigma^{-A} \vdash \Delta^{-A}, \Theta}\,(\text{mix})}{\Box\Gamma, \Box(\Sigma^{-A}) \vdash \Box(\Delta^{-A}), \Box\Theta}\,(\Box)$$

The sequences $\Box(\Sigma^{-A})$ and $(\Box\Sigma)^{-\Box A}$ are identical, as are $\Box(\Delta^{-A})$ and $(\Box\Delta)^{-\Box A}$, so the final lines of the two derivations match.

There is no case in which the $(\Box)$ rule is followed by a cut on a sentence that was parametric in $(\Box)$, as all sentences in the sequent are principal in that rule.

We conclude that cut is an admissible rule for LRT. □

As expected, LRT is equivalent to RT in the following sense.

Theorem 12. *1. If $RT \vdash A$ then the sequent $\vdash A$ is derivable in LRT*

2. If $A_1, \ldots, A_n \vdash B_1, \ldots, B_m$ is derivable in LRT, then $RT \vdash (A_1 \,\&\, \ldots \,\&\, A_n) \supset (B_1 \vee \ldots \vee B_m)$

Proof. The proof of (1) is straightforward. I will provide derivations of the K axiom, since the axioms of classical logic are derivable and the axioms of RT were shown derivable above. Since cut is admissible, we can show that *modus ponens* is admissible and the Nec rule is an instance of the $\Box$ rule.

$$\dfrac{\dfrac{\dfrac{A \vdash A \qquad B \vdash B}{A \supset B, A \vdash B}\,(\supset\vdash)}{\Box(A \supset B), \Box A \vdash \Box B}\,(\Box)}{\Box(A \supset B) \vdash \Box A \supset \Box B}\,(\vdash\supset)$$

The proof of (2) proceeds via an induction on the length of the proof. Since LRT is based on a classical multiple conclusion sequent calculus, I will omit the classical connective and structural rule steps. I will do the modal rule step here.

Suppose the last step of a proof is

$$\dfrac{A_1, \ldots, A_n \vdash B_1, \ldots, B_m}{\Box A_1, \ldots, \Box A_n \vdash \Box B_1, \ldots, \Box B_m}\,(\Box)$$

The RT proof proceeds as follows.

$(A_1 \,\&\, \ldots \,\&\, A_n) \supset (B_1 \vee \ldots \vee B_m)$	IH
$\Box((A_1 \,\&\, \ldots \,\&\, A_n) \supset (B_1 \vee \ldots \vee B_m))$	Nec
$\Box(A_1 \,\&\, \ldots \,\&\, A_n) \supset \Box(B_1 \vee \ldots \vee B_m)$	K axiom, *modus ponens*
$(\Box A_1 \,\&\, \ldots \,\&\, \Box A_n) \supset \Box(B_1 \vee \ldots \vee B_m)$	lemma 55, transitivity
$(\Box A_1 \,\&\, \ldots \,\&\, \Box A_n) \supset (\Box B_1 \vee \ldots \vee \Box B_m)$	lemma 55, transitivity

The final line is the desired conclusion.

□

In RT, $\Diamond$ is defined as $\sim\Box\sim$. We can take $\Diamond$ as a primitive along with $\Box$ and use the following, alternative rule.

$$\frac{A_1, \ldots, A_n \vdash B_1, \ldots, B_m}{\#_1 A_1, \ldots, \#_n A_n \vdash \#_{n+1} B_1, \ldots, \#_{n+m} B_m}\ (\Box\Diamond)$$

In this rule, each $\#_i$ is one of $\Box$ or $\Diamond$. To obtain the negation-free fragment, add the rules for the material conditional. Let RT^+ be the system RT with $(\Box\Diamond)$ replacing $(\Box)$. We have the following.

Theorem 13. *Cut is admissible for RT^+.*

RT^+ yields the desired relationships between $\Box$ and $\Diamond$.

Lemma 56. *The following are derivable in RT^+.*

1. $\Box A \vdash \sim\Diamond\sim A$
2. $\sim\Diamond\sim A \vdash \Box A$
3. $\Diamond A \vdash \sim\Box\sim A$
4. $\sim\Box\sim A \vdash \Diamond A$

Proof. I will derive only the first, as derivations for the others are basically the same.

$$\dfrac{\dfrac{\dfrac{A \vdash A}{A, \sim A \vdash}\ (\sim\vdash)}{\Box A, \Diamond\sim A \vdash}\ (\Box\Diamond)}{\Box A \vdash \sim\Diamond\sim A}\ (\vdash\sim)$$

□

The equivalence between $\Box$ and $\sim\Diamond\sim$, and between $\Diamond$ and $\sim\Box\sim$, extends to cases in which these connectives are embedded, as well.

Finally, we connect RT and RT^+.

Theorem 14. *Let $\Gamma \vdash \Delta$ be a sequent, possibly containing $\Diamond$, and let $\Gamma' \vdash \Delta'$ be the sequent obtained by replacing all instances of $\Diamond$ in Γ and Δ with $\sim\Box\sim$.*

(1) If the sequent $\Gamma \vdash \Delta$ is derivable in RT, $\Gamma \vdash \Delta$ is derivable in RT^+.

(2) If the sequent $\Gamma \vdash \Delta$ is derivable in RT^+, $\Gamma' \vdash \Delta'$ is derivable in RT.

Proof. (1) holds since $(\Box)$ is a special case of $(\Box\Diamond)$.

For (2), assume $\Gamma \vdash \Delta$ is derivable in RT^+. We describe how to modify the derivation to obtain a proof of $\Gamma' \vdash \Delta'$. The $\Diamond$s in $\Gamma \vdash \Delta$ must have come from either weakening or $(\Box\Diamond)$, since RT^+ is not formulated with cut.

If a $\Diamond$ came from weakening, then we replace the weakening step with one that weakens in A', replacing all $\Diamond$s in A with $\sim\Box\sim$.

Suppose that $\Diamond A$ came via $(\Box\Diamond)$, and suppose that $\Diamond A$ occurs on the left of the turnstile. We will do the case in which multiple $\Diamond$s are introduced, all of which are displayed. The case for $\Diamond A$ introduced on the right or on both sides at once is similar.

$$\frac{A_1, \ldots, A_n, \Sigma \vdash \Theta}{\Diamond A_1, \ldots, \Diamond A_n, \#\Sigma \vdash \#\Theta}\,(\Box\Diamond)$$

These steps will be replaced by the following.

$$\dfrac{\dfrac{\dfrac{A_1, \ldots, A_n, \Sigma \vdash \Theta}{\Sigma \vdash \sim A_1, \ldots, \sim A_n, \Theta}\,(\vdash\sim)}{\#\Sigma \vdash \Box\sim A_1, \ldots, \Box\sim A_n, \#\Theta}\,(\Box\Diamond)}{\sim\Box\sim A_1, \ldots, \sim\Box\sim A_n, \#\Sigma \vdash \#\Theta}\,(\sim\vdash)$$

The displayed application of $(\Box\Diamond)$ introduces only $\Box$s.

Removing all instances of $\Diamond$ in the RT^+ derivation using the preceding two steps results in an RT proof of $\Gamma' \vdash \Delta'$, as desired.

□

We can use RT^+ to study the absolutely positive fragment of RT, which is obtained by adding primitive $\Diamond$ and dropping $\sim$ and $\supset$.[3] We will not, however, investigate the absolutely positive fragment further here.

Before moving on to the quantified modal logic, we will note one more fact about LRT. The system LRT lacks a desirable feature: it does not uniquely define the connective $\Box$.[4] If another connective, $\circ$, is added with the same rules, the sequents $\Box A \vdash \circ A$ and $\circ A \vdash \Box A$ will not be derivable.

To obtain the quantified version of RT, RTQ, we add to RT the following axiom schemes and the rule of generalization: from $\vdash A$ infer $\vdash \forall x A$.

Q1 $\forall x(A \supset B) \supset \forall x A \supset \forall x B$

Q2 $\forall x A(x) \supset A(t)$, where t is free for x

Q3 $A \supset \forall x A$, where x is not free in A

Q4 $\forall x \Box A \supset \Box \forall x A$

The resulting system, as expected, proves the converse Barcan formula. The Barcan formula must be added as an axiom scheme. Perhaps unsurprisingly, in light of the failure of the Hilbert system to prove the Barcan formula, the result of adding the quantifier rules to the sequent system does not permit us to prove the Barcan formula, although it does permit a proof of the converse Barcan formula. We will illustrate the problem.

$$\dfrac{\dfrac{\dfrac{\dfrac{A \vdash A}{A \vdash \forall x A}\,(??)}{\Box A \vdash \Box \forall x A}\,(\Box)}{\forall x \Box A \vdash \Box \forall x A}\,(\forall \vdash)}$$

The proof breaks down at the step labelled (??). In order to apply the $(\vdash \forall)$ rule, x cannot be free in A. We are considering arbitrary A here, so the restriction appears to block the proof. We have not been able to adjust the sequent system so that it proves the Barcan formula. We have, however, formulated a hypersequent system in which the Barcan formula is provable. Before moving to the hypersequent system we will show how to derive the converse Barcan formula.

[3] The term "absolutely positive fragment" comes from Dunn (1995).

[4] The significance of this feature was first pointed out in Belnap (1962). I am indebted to Greg Restall for suggesting that I check it in this case.

$$\dfrac{\dfrac{\dfrac{A \vdash A}{\forall x A \vdash A}\,((\forall\vdash))}{\Box\forall x A \vdash \Box A}\,(\Box)}{\Box\forall x A \vdash \forall x \Box A}\,(\vdash\forall)$$

A *hypersequent* is a sequence of sequents, $\Gamma_0 \vdash \Delta_0 | \ldots | \Gamma_n \vdash \Delta_n$. The sequents between the vertical lines are the *components*. A hypersequent with n components is an n-hypersequent. We will use the notation $\mathcal{H}[\Gamma \vdash \Delta]$ for a hypersequent in which there is a particular occurrence of the component $\Gamma \vdash \Delta$.[5] We will use the notation $\mathcal{H}_n[\Gamma_k \vdash \Delta_k]$ to indicate that the hypersequent is an n-hypersequent and the displayed component is the kth one. We will also use the notation $\mathcal{H}[\Gamma \vdash_k \Delta]$ to indicate that the displayed component is the kth one. The notation $\mathcal{H}[\Gamma \vdash \Delta | \Gamma' \vdash \Delta']$ will be used for a hypersequent with the two displayed components, which are next to each other in that order. We will use $\emptyset_n$ for the n-hypersequent with all its components empty, with the convention that if $n \leq 0$, then $\emptyset_n$ is empty. We will abuse notation slightly and use $\Gamma \vdash \Delta \mid \mathcal{H}$ for the hypersequent obtained by prefixing the n-hypersequent $\mathcal{H}$ with the sequent $\Gamma \vdash \Delta$ to obtain an $n+1$-hypersequent.

The rules for the system $HRTQ$ are as follows. For each n, an axiom n-hypersequent is an n-hypersequent all of whose components are empty except for one, which has an axiom sequent.

$$\vdash | \ldots | A \vdash A | \ldots | \vdash$$

All of the classical connectives have the standard rules for each component.[6] I will give an example of one pair of rules. In the following, $\Sigma_i = \Gamma_i \cup \Gamma'_i$ and $\Theta_i = \Delta_i \cup \Delta'_i$.

$$\dfrac{\Gamma'_0 \vdash \Delta'_0 | \ldots | \Gamma'_k \vdash \Delta'_k, A | \ldots | \Gamma'_n \vdash \Delta'_n \quad \Gamma_0 \vdash \Delta_0 | \ldots | \Gamma_k \vdash \Delta_k, B | \ldots | \Gamma_n \vdash \Delta_n}{\Sigma_0 \vdash \Theta_0 | \ldots | \Sigma_k \vdash \Theta_k, A \,\&\, B | \ldots | \Sigma_n \vdash \Theta_n}\,(\vdash\&)$$

$$\dfrac{\mathcal{H}[\Gamma, A \vdash \Delta]}{\mathcal{H}[\Gamma, A \,\&\, B \vdash \Delta]}\,(\&\vdash) \qquad \dfrac{\mathcal{H}[\Gamma, B \vdash \Delta]}{\mathcal{H}[\Gamma, A \,\&\, B \vdash \Delta]}\,(\&\vdash)$$

Each component has the full set of structural rules, and there are no structural rules for manipulating the hypersequents. The quantifier rules are as follows.

$$\dfrac{\mathcal{H}[A, \Gamma \vdash \Delta]}{\mathcal{H}[\forall x A, \Gamma \vdash \Delta]}\,(\forall\vdash) \qquad \dfrac{\mathcal{H}[\Gamma \vdash \Delta, A]}{\mathcal{H}[\Gamma \vdash \Delta, \forall x A]}\,(\vdash\forall)$$

[5] Hypersequents were discovered by Pottinger and, independently, by Avron, in Pottinger (1983) and Avron (1987), respectively. Our notation is based on Restall (2012).

[6] See Restall (2012) for examples.

In the ($\vdash\forall$) rule, x cannot occur freely in the conclusion hypersequent. The modal rules are the following.

$$\frac{\mathcal{H}[\Gamma \vdash \Delta|\ A, \Gamma' \vdash \Delta']}{\mathcal{H}[\Box A, \Gamma \vdash \Delta|\ \Gamma' \vdash \Delta']}\ (\Box\vdash) \qquad \frac{\mathcal{H}[\Gamma \vdash \Delta|\ \Gamma' \vdash \Delta', A]}{\mathcal{H}[\Gamma \vdash \Delta, \Box A|\ \Gamma' \vdash \Delta']}\ (\vdash\Box)$$

The cut rule for $HRTQ$ is the following.

$$\frac{\Gamma'_0 \vdash \Delta'_0| \ldots |\, \Gamma'_k \vdash \Delta'_k, A| \ldots |\, \Gamma'_n \vdash \Delta'_n \quad \Gamma_0 \vdash \Delta_0| \ldots |A, \Gamma_k \vdash \Delta_k| \ldots |\, \Gamma_n \vdash \Delta_n}{\Sigma_0 \vdash \Theta_0| \ldots |\Sigma_k \vdash \Theta_k| \ldots |\Sigma_n \vdash \Theta_n}\ (\text{cut})$$

When the contexts are the same, we can rewrite this in a simpler way.

$$\frac{\mathcal{H}_n[\Gamma \vdash_k \Delta, A] \quad \mathcal{H}_n[A, \Gamma \vdash_k \Delta]}{\mathcal{H}_n[\Gamma \vdash_k \Delta]}$$

The system $HRTQ$ does not have the cut rule among its basic rules, but we can show the following.

Theorem 15 (Cut elimination for $HRTQ$)**.** *The cut rule is admissible for HRTQ*

Proof. The proof proceeds via a double induction on the depth and rank of the cut formula. As in the proof of cut elimination earlier in this chapter, the proof uses the stronger rule of mix, rather than cut. The main change to the proof is that the rank of a formula is relative to a component. Note that the only the box rules move formulas between components. The rules and axioms within a given component are all classical, so I will omit many of the steps. The important new step is the step in which the cut formula is $\Box A$ and the cut formula was just introduced on both sides. I will use the notation $\mathcal{H}'$ for the wider context of $\Gamma'_i \vdash \Delta'_i$, for $i \neq k, k+1$ and $\mathcal{H}''$ for wider context of $\Sigma_i \vdash \Theta_i$, for all $i \neq k, k+1$.

$$\dfrac{\dfrac{\mathcal{H}'[\Gamma'_k \vdash \Delta'_k|\ \Gamma'_{k+1} \vdash \Delta'_{k+1}, A]}{\mathcal{H}'[\Gamma'_k \vdash \Delta'_k, \Box A|\ \Gamma'_{k+1} \vdash \Delta'_{k+1}]}\ (\vdash\Box) \quad \dfrac{\mathcal{H}[\Gamma_k \vdash \Delta_k|\ A, \Gamma_{k+1} \vdash \Delta_{k+1}]}{\mathcal{H}[\Box A, \Gamma_k \vdash \Delta_k|\ \Gamma_{k+1} \vdash \Delta_{k+1}]}\ (\Box\vdash)}{\mathcal{H}''[\Sigma_k \vdash \Theta_k | \Sigma_{k+1} \vdash \Theta_{k+1}]}\ (\text{mix})$$

The mix can be carried out on simpler formulas in the following way.

$$\dfrac{\dfrac{\mathcal{H}'[\Gamma'_k \vdash \Delta'_k|\ \Gamma'_{k+1} \vdash \Delta'_{k+1}, A] \quad \mathcal{H}[\Gamma_k \vdash \Delta_k|\ A, \Gamma_{k+1} \vdash \Delta_{k+1}]}{\mathcal{H}''[\Sigma_k \vdash \Theta_k \mid \Gamma'_{k+1}, \Gamma^{-A}_{k+1} \vdash \Delta_{k+1}, \Delta'^{-A}_{k+1}\ ]}\ (\text{mix})}{\mathcal{H}''[\Sigma_k \vdash \Theta_k \mid \Gamma'_{k+1}, \Gamma_{k+1} \vdash \Delta_{k+1}, \Delta'_{k+1}\ ]}\ (K)$$

□

We will note three interesting features of $HRTQ$. First, weakening-style rules for hypersequents are admissible.

Theorem 16. *The rules*

$$\frac{\mathcal{H}}{\Gamma \vdash \Delta \mid \mathcal{H}} \qquad \frac{\mathcal{H}}{\mathcal{H} \mid \Gamma \vdash \Delta}$$

are admissible, where $\mathcal{H}$ is a hypersequent.

Proof. I will sketch the proof for one, since the proof for the other is similar. The proof proceeds by induction on the length of the proof. For a given proof of $\mathcal{H}_n$, replace all axioms in the proof with the corresponding $n+1$-hypersequent axiom, with the position of the formulas determined by counting from the right. The proof can be carried out using the same inference rules as before to obtain $\vdash \mid \mathcal{H}$. Repeated instances of the K rule can then be used to obtain the desired end hypersequent. □

This immediately gives the following two corollaries. Let us say that if $\mathcal{H}_n[\vdash_0 A]$ is derivable where the wider context is empty, then A is an *n-theorem*. The following corollary will be used in an equivalence proof below.

Corollary 13. *If A is an n-theorem, then, A is an $n+m$-theorem, for all $m \geq 0$.*

Corollary 14. *The rules*

$$\frac{\mathcal{H}}{\vdash \mid \mathcal{H}} \qquad \frac{\mathcal{H}}{\mathcal{H} \mid \vdash}$$

are admissible and height-preserving.

Proof. Admissibility follows from theorem 16. Height-preservation follows from the fact that we only change the number of components in the axioms, not any of the rules used.

□

The second feature of $HRTQ$ is that the identity theorem fails for it, since $\Box A \vdash \Box A$ is not derivable for arbitrary formulas A. Note, however, that a slightly weakened form is derivable, where $d(A)$ is the modal depth of A.

Theorem 17. *If $d(A) \leq n$, then $A \vdash A \mid \emptyset_n$ is derivable.*

Proof. The proof is by a double induction on the modal depth and complexity of A.

If $d(A) = 0$, then $A \vdash A$ is derivable, because the identity theorem holds for classical logic.

Assume that for all formulas B such that $d(B) \leq n$, $B \vdash B|\emptyset_{n-1}$ is derivable. We now proceed by induction on the complexity of A.

If A is of the form $\Box B$, then we have the following derivation, whose first line is justified by the assumption whose second is justified by the previous corollary.

$$\dfrac{\dfrac{\dfrac{B \vdash B \mid \emptyset_{n-1}}{\vdash \mid B \vdash B \mid \emptyset_{n-1}}\,(\vdash\Box)}{\vdash \Box B \mid B \vdash \mid \emptyset_{n-1}}\,(\Box\vdash)}{\Box B \vdash \Box B \mid \vdash \mid \emptyset_{n-1}}$$

We note that if $d(B) \leq n$, then by the induction hypothesis and corollary 13, there is a derivation of $B \vdash B|\emptyset_n$. The remaining cases are taken care of by the induction hypothesis on complexity, with the following being used in binary connective subcases, $B \circ C$, in which either $d(B) \leq n$ or $d(C) \leq n$.

□

A third interesting feature of $HRTQ$, which is the main reason it is being discussed, is that it permits derivations of the instances of the Barcan formula, unlike LRT with the quantifier rules. Here is a proof of the Barcan formula, given a proof of $A \vdash A \mid \emptyset_{n-1}$.

$$\dfrac{\dfrac{\dfrac{\dfrac{\vdash \mid A \vdash A \mid \emptyset_{n-1}}{\Box A \vdash \mid \vdash A \mid \emptyset_{n-1}}\,(\Box\vdash)}{\forall x \Box A \vdash \mid \vdash A \mid \emptyset_{n-1}}\,(\forall\vdash)}{\forall x \Box A \vdash \mid \vdash \forall x A \mid \emptyset_{n-1}}\,(\vdash\forall)}{\forall x \Box A \vdash \Box \forall x A \mid \emptyset_n}\,(\vdash\Box)$$

$HRTQ$ is equivalent to RTQ in the following way. Define a function f on a hypersequent $\Gamma_0 \vdash \Delta_0 | \ldots | \Gamma_n \vdash \Delta_n$ as

$$f(\Gamma_0 \vdash \Delta_0 | \ldots | \Gamma_n \vdash \Delta_n) = (\bigwedge \Gamma_0 \supset \bigvee \Delta_0) \vee \Box(f(\Gamma_1 \vdash \Delta_1 | \ldots | \Gamma_n \vdash \Delta_n)).^{7}$$

If $X = \emptyset$, then $\bigwedge X = \top$, and $\bigvee X = \bot$.

[7] If $\Gamma = A_1, \ldots, A_n$, then $\bigwedge \Gamma = A_1 \,\&\, \ldots \,\&\, A_n$ and $\bigvee \Gamma = A_1 \vee \ldots \vee A_n$.

Before stating and proving the equivalence theorems, I will prove one required lemma concerning an unusual rule in RT.

Lemma 57. *The sentence*

$$(X \vee (A \supset B) \vee \Box(C \,\&\, D \supset E)) \supset (X \vee (A \,\&\, \Box C \supset B) \vee \Box(D \supset E))$$

is valid in RT.

Proof. Suppose (M, w) is a pair of an RT model and world such that

$$M, w \Vdash X \vee (A \supset B) \vee \Box(C \,\&\, D \supset E)$$

and

$$M, w \nVdash X \vee (A \,\&\, \Box C \supset B) \vee \Box(D \supset E).$$

Then, $M, w \nVdash X$, $M, w \nVdash (A \,\&\, \Box C \supset B)$ and $M, w \nVdash \Box(D \supset E)$. Suppose that wRy, so y is the unique successor of w. $M, w \Vdash A \,\&\, \Box C$, so $M, w \Vdash A$ and $M, y \Vdash C$. By assumption, $M, v \nVdash B$. $M, w \nVdash \Box(D \supset E)$, so $M, y \Vdash D$ and $M, y \nVdash E$. By assumption, $M, w \Vdash X \vee (A \supset B) \vee \Box(C \,\&\, D \supset E)$, so either $M, w \Vdash X$, $M, w \Vdash A \supset B$, or $M, w \Vdash \Box(C \,\&\, D \supset E)$.

Case 1: $M, w \Vdash A \supset B$. Since $M, w \Vdash A$, $M, w \Vdash B$, which is a contradiction.

Case 2: $M, w \Vdash \Box(C \,\&\, D \supset E)$. So $M, y \Vdash C \,\&\, D \supset E$. As $M, y \Vdash C \,\&\, D$, $M, y \Vdash E$, which is a contradiction.

Case 3: $M, w \Vdash X$. The contradiction is immediate.

In all cases we reached a contradiction, so we conclude by *reductio* that

$$M, w \Vdash (X \vee (A \supset B) \vee \Box(C \,\&\, D \supset E)) \supset (X \vee (A \,\&\, \Box C \supset B) \vee \Box(D \supset E)).$$

□

We can now prove the equivalence of RTQ and $HRTQ$

Theorem 18. *1. If $RTQ \vdash A$, then for some n, A is an n-theorem of $HRTQ$.*

2. If $\mathcal{H}$ is derivable in $HRTQ$, then $RTQ \vdash f(\mathcal{H})$.

Proof. The proof of (1) proceeds by showing that all axioms are derivable and that if the premises of the rules are derivable, then so are the conclusions. The main difficulty of the proof is determining values of n. Given a theorem A of RT, we will use the depth of the greatest nesting of boxes in A for the value of n. We have $A \vdash A|\emptyset_{n-1}$, by the identity theorem. We can then obtain the following.

$$\dfrac{\dfrac{\dfrac{\vdash \mid \vdash A, \sim A \mid \emptyset_{n-1}}{\vdash \Box A, \Box \sim A \mid \emptyset_n}\ (\vdash \Box)}{\sim \Box A \vdash \Box \sim A \mid \emptyset_n}\ (\sim \vdash)}{\vdash \sim \Box A \supset \Box \sim A \mid \emptyset_n}\ (\vdash \supset)$$

The derivations of the other propositional axioms are similar. I provided a derivation for the Barcan formula. The other quantifier axioms are similar. There are three rules remaining to prove: *modus ponens*, generalization and necessitation.

Generalization follows from the $\vdash \forall$ rule. Necessitation follows from the weakening rule and the $\vdash \Box$ rule. The last rule, *modus ponens*, follows from corollary 13. Suppose that A is n-theorem and $A \supset B$ is an m-theorem. Let $k = max(n, m)$. Then, by corollary 13, both are k-theorems. There is a proof of

$$\vdash A \supset B \mid \emptyset_{k-1},$$

which is obtained via the $\vdash \supset$ rule, which makes the penultimate hypersequent

$$A \vdash B \mid \emptyset_{k-1}.$$

By the elimination theorem, there is then a proof of

$$\vdash B \mid \emptyset_{k-1}$$

that does not use cut. This is the desired conclusion.

The proof of (2) proceeds by inductively showing that the translations of the axioms of HRTQ are derivable in RTQ and that if the translations of the premises of a rule are derivable in RTQ then so is the translation of the conclusion. There are many cases to

check, so I will present sketch of the more interesting ones here. First, an axiom case. The axioms are hypersequents of the form

$$\mathcal{H}[A \vdash A],$$

where the non-displayed components are all empty. Suppose that there are n components and the displayed one is the kth component. The translation of this hypersequent is equivalent to $\Box^k(A \supset A)$, which is a theorem of RTQ.

For the $\Box\vdash$ case, assume that $\mathcal{H}[\Gamma_k \vdash \Delta_k \mid \Gamma_{k+1}, A \vdash \Delta_{k+1}]$ is derivable in HRTQ and that the translation is derivable in RTQ. We want to show that the translation of $\mathcal{H}[\Gamma_k, \Box A \vdash \Delta_k \mid \Gamma_{k+1} \vdash \Delta_{k+1}]$ is derivable in RTQ. This case is an instance of lemma 57, with extra disjunctions. By assumption the translation of the premiss is derivable in RTQ, so the conclusion is derivable by lemma 57.

For the $\vdash\forall$ case, assume that $\mathcal{H}[\Gamma_k \vdash \Delta_k, A]$ is derivable in $HRTQ$ and that the translation is derivable in RTQ. We want to show that the translation of $\mathcal{H}[\Gamma_k \vdash \Delta_k, \forall x A]$, where x is not free in $\mathcal{H}$, is derivable in RTQ. The translation of the premiss is equivalent to $X \vee \Box^k(\bigwedge \Gamma_k \supset \bigvee \Delta_k \vee A) \vee Y$, for some formulas X and Y. By assumption this is derivable in RTQ. By the rule of generalization,

$$\forall x(X \vee \Box^k(\bigwedge \Gamma_k \supset \bigvee \Delta_k \vee A) \vee Y)$$

is derivable in RTQ. Since x is not free in $\mathcal{H}[\Gamma_k \vdash \Delta_k, \forall x A]$, x is not free in $\mathcal{H}[\Gamma_k \vdash \Delta_k, A]$, except possibly in the displayed occurrence of A. Therefore, by use of the Barcan formula and the logical truth $\forall x(A \vee B) \supset (A \vee \forall x B)$, where x is not free in A, $\forall x(X \vee \Box^k(\bigwedge \Gamma_k \supset \bigvee \Delta_k \vee A) \vee Y)$ implies

$$X \vee \Box^k \forall x(\bigwedge \Gamma_k \supset \bigvee \Delta_k \vee A) \vee Y.$$

From this we obtain

$$X \vee \Box^k(\bigwedge \Gamma_k \supset \bigvee \Delta_k \vee \forall x A) \vee Y,$$

which was to be shown.

Contraction is handled by the idempotence of & and $\vee$, depending on whether the contraction is on the left or the right. The other structural rules are also straightforward. $\Box$

Before we move on from $HRTQ$ we will note another hypersequent system that may be worth investigating. As it stands, it is hard to see how to prove completeness for $HRTQ$ directly, since each axiom permits only finitely many components. An alternative would be to formulate the system using a $\mathbb{Z}$-chain of components with a designated component, zero. If all components above or below a given component are empty, one could omit them from the notation, so the system would notationally be similar to $HRTQ$. The upshot is that it would, we hope, be possible to prove a completeness theorem, using methods similar to those of Restall (2009).[8] We will, however, set that idea aside and return to RT and RTQ.

For the next theorems, we will assume that the primitive non-modal connectives are: $\&, \sim$, and, in the first-order case, $\forall$. We now define a reduction relation, $\rhd$, on pairs of sentences (formulas) in the language of $RT(RTQ)$. Roughly, $A \rhd^* A'$ iff A' is the result of simultaneously pushing all boxes in A one level in.

First, let us define two functions for use in the definition of $\rhd^*$.

$$A^\dagger = \begin{cases} p & \text{if } A \text{ is } p \\ B^\dagger \,\&\, C^\dagger & \text{if } A \text{ is } B \,\&\, C \\ \sim(B^\dagger) & \text{if } A \text{ is } \sim B \\ \forall x(B^\dagger) & \text{if } A \text{ is } \forall x B \\ (\Box B)^\circ & \text{if } A \text{ is } \Box B \end{cases} \qquad A^\circ = \begin{cases} \Box p & \text{if } A \text{ is } \Box p \\ \Box(B^\dagger) \,\&\, \Box(C^\dagger) & \text{if } A \text{ is } \Box(B \,\&\, C) \\ \sim\Box(B^\dagger) & \text{if } A \text{ is } \Box\sim B \\ \forall x\Box(B^\dagger) & \text{if } A \text{ is } \Box\forall x B \\ \Box(\Box B)^\circ & \text{if } A \text{ is } \Box\Box B \end{cases}$$

Next, $A \rhd^* A'$ iff $A' = A^\dagger$. Finally, let $\rhd$ be the reflexive, transitive closure of $\rhd^*$.

Let us define *box normal form*, or bnf, of A as a wff B such that $A \rhd B$ and for all C, if $B \rhd C$, then $B = C$. Every formula has a bnf, which is unique.

There are two theorems concerning RT and RTQ that I will now prove. For these proofs, I will use the notation A' for the bnf of A. I will also use *tautology*, and *first-order logical truth*, to mean, respectively, a substitution instance of a classical tautology, or classical first-order logical truth, in the language of RT, or RTQ.

Theorem 19. *For all sentences A, $RT \vdash A$ iff A' is a tautology.*

[8] The connection was inspired by Restall's presentation, "Exotic Sequent Calculi for Truth Degrees," at the Logic, Algebra and Truth Degrees 2012 conference.

Proof. The right-to-left direction follows since RT entails that the box commutes with all connectives and RT entails all tautologies.

The proof of the converse is by induction on the proof of A.

The base cases are the axioms.

First the tautologies. The tautologies can be taken to be axiomatized by the standard axioms for $\supset$ and $\sim$. The result of replacing A and B in $A \supset (B \supset A)$ with their respective bnfs is an instance of the axiom, and similarly for the other axioms.

For the special RT axiom, I will use the form $\sim\Box A \equiv \Box\sim A$. This is equivalent in RT to $\sim\Box A \equiv \sim\Box A$. The result of replacing $\Box A$ with its bnf, A', yields $\sim A' \equiv \sim A'$.

The modal axiom $\Box(A \supset B) \supset .\Box A \supset \Box B$ is, in RT, equivalent to $\Box A \supset \Box B \supset .\Box A \supset \Box B$. The result of replacing $\Box A$ and $\Box B$ with their respective bnfs A' and B' is a tautology.

For the inductive steps, we need to show that the rules *modus ponens* and Necessitation preserve the desired property.

Inductive case: Necessitation. Suppose that $RT \vdash A$. Then, by the induction hypothesis, A has a bnf A' that is a tautology. By Necessitation, $RT \vdash \Box A$. The result of adding an additional box to each maximal boxed atom of A' is a tautology, which completes this step.

Inductive case: *modus ponens*. Suppose that $RT \vdash A$ and $RT \vdash A \supset B$. By the induction hypothesis, there are bnfs A' and C' of A and $A \supset B$, respectively, such that A' and C' are tautologies. From the definition of bnf, C is $A' \supset B'$, where B' is the bnf of B. Tautologies are closed under *modus ponens*, so B' is a tautology, as desired.

We conclude that for all A, $RT \vdash A$ iff A' is a tautology.

□

Theorem 20. *For all formulas A, $RTQ \vdash A$ iff A' is a first-order logical truth.*

Proof. The proof of this is by induction on the proof of A. It is largely similar to the preceding proof, with additional steps for the quantifier axioms, the Barcan formula, and the rule of Generalization. These additional steps are straightforward, so we omit them.

□

It is worth observing that the preceding theorem does not hold if we include equality in the language. The reason is that RTQ with equality axioms proves $\Box^n x = x$, for all n, but

$\Box^n x = x$ is not a substitution instance of a first-order logical truth if $n \geq 1$.

The following definition will be useful in some of the later equivalence proofs.

Definition 24 (k-transform)**.** *For $k \geq n$, the* k-transform *of a formula A^n is $\Box^{k-n} A^k$.*

Our interest in k-transforms comes primarily from the following lemma, whose proof we omit.

Lemma 58. *Suppose that there is a deduction $A_1^{k_1}, \ldots, A_n^{k_n} \vdash_0^{\mathscr{D}} B^{k_{n+1}}$ of an indexed formula from a set of indexed formulas. Let $k = max(\{k_i : 1 \leq i \leq n+1\})$ Then, there is a deduction $A_1'^k, \ldots, A_n'^k \vdash_0^{\mathscr{D}} B'^k$, where B' is the k-transform of B and, for each i, A_i' is the k-transform of A_i.*

Let $RTQ^=$ be RTQ with the following equality axioms.[9]

E1 $t = t$

E2 $s = t \supset (A \supset A')$, where A and A' differ only in that A' has t in 0 or more places that A has s, assuming t is free for s in A

E3 $s \neq t \supset \Box(s \neq t)$

Given the other axioms of $RTQ^=$, E3 implies

$$\Box(s = t) \supset s = t.$$

Let C^- be the Fitch system $C_0^{\Box}$ without the definition rules and with the index shift restricted to just identity statements.

Theorem 21 (Equivalence)**.** *The following are equivalent.*

1. *C^- proves A^0.*
2. *$RTQ^=$ proves A*

Proof. The proof of (2)⇒(1) is straightforward. In the completeness section we demonstrated that most of the axiom schemes of $RTQ^=$ were derivable in C^-. We will note that index shift for identity statements is sufficient, and seems to be necessary, for deriving E3. We will omit the proofs of the axioms.

[9] These are taken from Hughes and Cresswell (1996).

The proof of (1) $\Rightarrow$ (2) is more involved. Suppose we are given a derivation C^- derivation of A^0. Let k be the greatest index appearing in the derivation. Replace each line of the derivation with its k-transform. The result is an admissible proof. We now proceed inductively to replace innermost subproofs with the "quasi-proof" method of Anderson and Belnap (1975).

There are many cases, and they are mostly straightforward. We will merely indicate how to fill in two of the more interesting cases.

Case: $\forall$I. This case is handled by repeated use of substitution of equivalents and appropriate instances of the Barcan formula.

Case: IS rule. $RTQ^= \vdash t = s \equiv \Box(t = s)$, and the k-transform of the conclusion of an index shift rule will be one direction of that biconditional.

□

The preceding work shows two things. First, the modal logics RT and RTQ are both simple. In a way, RT is simpler than the modal logic K.[10] Second, the modal logic $RTQ^=$ has a clear connection with C^-. The relationship between the modal logics and the revision theory will be further explored in the next section.

We will close this section by noting a problem we are leaving open. The formulation of $HRTQ$ does not include equality rules. It would be good to be able to add equality to the system in a way that permits either an elimination theorem or an elimination theorem with cuts only on atoms. One possible application that is suggested by the next section is to use ideas from $HRTQ$ to develop a hypersequent calculus for finite definitions that does not use indices.

[10]The simplicity of RT does not carry over to all its presentations. We can formulate a display calculus for it, and its quantified form, using the methods of Wansing (1998). For this, make Wansing's structure for $\Diamond$, $* \bullet *$ display equivalent to the structure for $\Box$, $\bullet$, when either both are in antecedent position or both in consequent position. For more on display calculuses, see Belnap (1982a).

5.2 SOLOVAY-TYPE THEOREMS

There is a connection between the modal logic RT, provability in $C_0^{\Box}$, and validity for finite definitions. This connection is similar to Solovay's arithmetical completeness theorem relating the provability logic GL and provability in Peano arithmetic.[11] The connection is made via *arithmetical interpretations*. An arithmetical interpretation * for a propositional modal language is a function that maps sentences of the modal language to sentences of PA. An interpretation is defined as follows.

- $(p)^* \in Sent_{\mathscr{L}_{PA}}$
- $(\bot)^* = \bot$
- $(\sim A)^* = \sim(A^*)$
- $(A \supset B)^* = (A^*) \supset (B^*)$
- $(\Box A)^* = Pr(\ulcorner A^* \urcorner)$

The conceptually most important clause is the one stating that the box is interpreted as a provability predicate.

Theorem 22 (Solovay's Theorem)**.** *$GL \vdash A$ iff for all arithmetical interpretations *, $PA \vdash A^*$.*

For a proof see Boolos (1993, 125-131). Different completeness theorems can be proved by restricting to different sets of sentences for the interpretation of atoms. First-order interpretations can be defined as well. Atomic formulas are mapped to formulas of PA with the same free variables and the interpretations commute with the quantifiers. The analogous first-order completeness theorem does not hold.[12] In this section we will prove similar completeness theorems for the logics RT and RTQ, using interpretations based on finite definitions rather than arithmetic sentences. We will begin by proving the theorem for the simplest case, propositional logic. This will permit us to clearly demonstrate the proof technique. We will then prove the theorem for the first-order logic and conclude by sketching some variations.

[11] Solovay had other results in this area. I state just one here. For more, see Boolos (1993), especially chapter 5.

[12] See Boolos (1993, Ch. 17-18) for details.

5.2.1 Propositional Solovay-type theorem

The first task is to define the relevant notion of interpretation. Given a base language $\mathscr{L}$, expand it to $\mathscr{L}^+_{\mathscr{D}}$ with the addition of $\Box$ and a set of definitions $\mathscr{D}$. Let a $\mathscr{D}$-interpretation * be a function from sentences of RT to sentence of $L^+_{\mathscr{D}}$ satisfying the following.

- $(p)^* \in Sentence_{\mathscr{L}^+_{\mathscr{D}}}$
- $(\bot)^* = \bot$
- $(\sim A)^* = \sim(A^*)$
- $(A \circ B)^* = (A^*) \circ (B^*)$, for $\circ \in \{\&, \vee, \supset\}$
- $(\Box A)^* = \Box(A^*)$

We are now ready to state the Solovay-like theorem relating RT, $C^{\Box}_0$ and simple finite definitions.

Theorem 23. *Let $\mathscr{L}$ be a ground language with at least one name.*

1. $RT \vdash A$ *iff* $\forall \mathscr{D}\, \forall \mathscr{D}$*-interpretations* $^*\, \forall k \vdash^{\mathscr{D}}_0 (A^*)^k$
2. $RT \vdash A$ *iff* $\forall$ *simple finite* $\mathscr{D}\, \forall \mathscr{D}$*-interpretations* $^*\, \models^{\mathscr{D}}_{\#} (A^*)$

I will briefly outline the proof strategy before providing the details. The left-to-right direction is proved by showing that the $\mathscr{D}$-interpretations of an axiom of RT are valid for finite definitions and that if the interpretation of the premises of *modus ponens* and necessitation are valid, so is the interpretation of the conclusion.

For the right-to-left direction, we argue contrapositively. Assume that $RT \nvdash A$, so $RT \nvdash A_{bnf}$, where A_{bnf} is the box normal form of A, in which all boxes are pushed in as far as possible. We use the completeness of RT to obtain a canonical model falsifying A_{bnf}. This gives a finite set of worlds and a distribution of truth values falsifying A_{bnf}. This is translated to a finite number of steps of a revision sequence, which allows us to determine the definition needed to produce that pattern of truth values. Finally, we show that the transformations used to obtain the box normal form are provably equivalent in $C^{\Box}_0$.

Figure 5 summarizes the proof strategy for the right-to-left direction. The double line arrows indicate the steps taken in the proof. The wavy arrow is the desired conclusion. The two steps represented by the vertical arrows on the left and right are accounted for

$$\begin{array}{ccc} RT \nvdash A & \rightsquigarrow & \nvdash_0^{\mathscr{D}} A^* \\ \Downarrow & & \Uparrow \\ RT \nvDash A_{bnf} & \Rightarrow & \nvDash_0^{\mathscr{D}} (A_{bnf})^* \end{array}$$

Figure 5: Proof strategy for Solovay-type theorems

by completeness and soundness results, respectively. Our primary contribution is the step represented by the arrow along the bottom, namely showing how to determine an appropriate $\mathscr{D}$ and define the invalidating model.

First, I will mention two useful results concerning RT.

Lemma 59. *For every A, $RT \vdash A \equiv A_{bnf}$.*

Proof. This follows from lemma 55. □

Theorem 24 (RT completeness)**.** *RT is sound and complete with respect to the class of Kripke frames in which every world has exactly one R successor.*

Proof. This is proved using the methods of Hughes and Cresswell (1996). □

We will make use of the following theorem, stated without proof.

Theorem 25. *Let M be an RT model, w a world of M, and k a natural number. Let $M'(= \langle W', R', V' \rangle)$ be the restriction of M to k-many R-successors of w, so that $w_0 R w_1 R, \ldots, R w_k$, with $w = w_0$. Then, for all $j \leq k$, for all A such that $d(A) \leq (k - j)$, $M, w_j \Vdash A$ iff $M', w_j \Vdash A$.*[13]

We will note that the first-order version of this holds for the constant domain models of RTQ.

Let A_{bnf} be the box normal form of A. Assume that $RT \nvdash A$, so $RT \nvdash A_{bnf}$. Then by the completeness theorem, there is a model M and a world w_0 such that $M, w_0 \nVdash A_{bnf}$. Let the modal depth of A be k. Then there are $k + 1$ worlds, such that $w_0 R w_1 \ldots R w_k$. For each

[13] Our theorem is specific to RT. See Blackburn et al. (2002, 76) for a more general version of this theorem.

maximal boxed atom, $\Box^n p$, of A, $M, w_0 \Vdash \Box^n p$ just in case $M, w_n \Vdash p$. We associate with each atom p, a $k+1$-long sequence of truth values, $\mathbf{p}$, such that $\mathbf{p}_i = \mathbf{t}$ just in case $M, w_i \Vdash p$, otherwise $\mathbf{p} = \mathbf{f}$. We will use the set of $\mathbf{p}$'s to define the required model and hypothesis.

Next, we will prove a lemma concerning definitions.

Lemma 60. *Let $\mathscr{G}$ be the set of definitions containing, for each n, m,*

$$G_n^m x =_{Df} \Box^n G_n^m x.$$

Then, every subset Y of $\mathscr{G}$ containing only finitely many definitions is finite.

Proof. We have to consider the behavior after revision of each subformula of the *definientia* in Y. The superscripts of the G_n^m do not affect the proof, so I will drop them. For any given G_n, the subformula of its definientia are $G_n x, \Box G_n x, \ldots, \Box^n G_n x$. Let h be an arbitrary hypothesis and let $\mathbf{X}_k$ be the extension h assigns to $\Box^k G_n x$. The extension of $G_n x$ after k revisions of h is $\mathbf{X}_p$, where $p = k \ mod \ (n+1)$. The extensions of each of the $\Box^k G_n x$ cycle with revisions. Hence, all hypotheses for $G_n a$ are reflexive.

The extension of each G_n cycles, so the extensions for the subformulas of finite sets of the definitions will cycle. The cycle will have the period of the least common multiple of the subscripts on all the G_n in the set of definitions. □

Only finitely many distinct atoms can appear in A_{bnf}, so a finite $\mathscr{D} \subseteq \mathscr{G}$ will suffice for the rest of the proof. In fact, the subset we will use is even simpler, since all the G's in the set of definitions will have the same subscript.

At this point, it may be helpful to see an example of a cycle of definitions from $\mathscr{G}$. Take $Gx =_{Df} \Box^3 Gx$ and some model M. The revisions of a particular hypothesis, h cycle in a simple pattern, as illustrated by table 1.[14] For definition G_n^m, the initial hypothesis will return to its values over the subformulas of the definition for G_n^m every $n+1$ revisions, with the pattern of satisfaction lagging on stage behind, as illustrated by table 2.

The *RT* countermodel provides a $k+1$-long sequence of R-successor worlds, starting with w_0. This provides a sequence of truth values, $\mathbf{p}$, for each atom p of A_{bnf}. The desired set of definitions is the set $G_k^1, \ldots, G_k^j$, where j is the number of distinct atoms in A. We can use

[14] The table uses truth values rather than pairs from hypotheses. The correspondence is straightforward.

	h	$\Delta(h)$	$\Delta^2(h)$	$\Delta^3(h)$	$\Delta^4(h)$	$\Delta^5(h)$	...
Ga	**t**	**t**	**f**	**f**	**t**	**t**	
$\Box Ga$	**f**	**t**	**t**	**f**	**f**	**t**	
$\Box^2 Ga$	**f**	**f**	**t**	**t**	**f**	**f**	
$\Box^3 Ga$	**t**	**f**	**f**	**t**	**t**	**f**	

Table 1: Pattern of truth values across revisions

any $\mathscr{D}$-interpretation such that sets $(p_i)^* = G_k^i a$, for each atom p_i of A. The interpretation of all other atoms can be arbitrary.

Next, we must construct the hypotheses to be used to falsify the interpretation of A_{bnf}. The model M can be arbitrary. For the hypotheses, we use the sequences $\mathbf{p}$ to determine the values for the G_k^i at each stage of revision. Define the sequence $\langle h_m \rangle_{m\in\omega}$ as follows.

- $\langle \Box^n G_n^i a, v\rangle \in_M h_0$ iff $M, w_n \Vdash p_i$, for each atom p_i.
- $h_{m+1} = \Delta_{M,\mathscr{D}}(h_m)$.

The sequence $\langle h_m \rangle_{m\in\omega}$ is a revision sequence for $\mathscr{D}$.

Next, we need to show that $M \not\models_0^{\mathscr{D}} A_{bnf}^*$. We use the sequence of hypotheses $\langle h_m \rangle_{m\in\omega}$. After k revisions, $M, h_k \not\models A_{bnf}^*$ For every p, $h_k \equiv_{\mathscr{D}} h_{p\cdot(k+1)+k}$, so, for all p, $M, h_{p\cdot(k+1)+k} \not\models A_{bnf}^*$ Therefore, $M \not\models_0^{\mathscr{D}} A_{bnf}^*$.

Corollary 15. *1.* $\not\models_0^{\mathscr{D}} A_{bnf}^*$

2. $\not\vdash_0^{\mathscr{D}} A_{bnf}^*$

3. $\not\models_{\#}^{\mathscr{D}} A_{bnf}^*$

To finish the proof, we must prove the following lemma.

Lemma 61. *For all $\mathscr{D}$ and all *, $\vdash_0^{\mathscr{D}} (A \equiv A_{bnf})^*$.*

This lemma is a corollary of the following.

Lemma 62. *$C_0^{\Box}$ proves each of the following.*

1. $((\Box\sim A) \equiv (\sim\Box A))^*$

$\models$	h	$\Delta(h)$	$\Delta^2(h)$	$\Delta^3(h)$	$\Delta^4(h)$	$\Delta^5(h)$	...
Ga	**t**	**f**	**f**	**t**	**t**	**f**	
$\Box Ga$	**t**	**t**	**f**	**f**	**t**	**t**	
$\Box^2 Ga$	**f**	**t**	**t**	**f**	**f**	**t**	
$\Box^3 Ga$	**f**	**f**	**t**	**t**	**f**	**f**	

Table 2: Pattern of satisfaction across revisions

2. $(\Box(A \,\&\, B) \equiv (\Box A \,\&\, \Box B))^*$

3. $((\Box(A \vee B) \equiv (\Box A \vee \Box B))^*$

4. $((\Box(A \supset B)) \equiv (\Box A \supset \Box B))^*$

Proof. $C_0^{\Box}$ has every instance of every equivalence schema, without the interpretations, as a theorem, so it has every interpretation of every instance of the schemas as a theorem. □

We can prove a variation of the theorem. Instead of using all simple finite definitions from the expanded revision theory with $\Box$, we can restrict to the set of definitions that are simple and finite in the original form of the revision theory. Rather than step through the proof in detail, I will highlight the changes that need to be made.

We will assume that the modal depth of A is at least 1, otherwise we are just dealing with classical validity.

We restrict to languages that have a binary relation symbol '$<$'. Let $LINORD(n)$ be a sentence saying that '$<$' is a discrete linear order with a least element, 0, and there are at least n distinct objects in the ordering. If there are not names for these n objects, enrich the language with them. Let $RESET(m, n)$ be a sentence saying that there are no more than n objects satisfying H_n^m and no object outside the ordering satisfies H_n^m. Let $\mathscr{H}$ be the set of the following definitions, for each $m, n \in \omega$.

$$H_n^m x =_{Df} LINORD(n) \,\&\, \sim RESET(m, n) \,\&\, [(\forall y(0 \leq y < x \supset H_n^m y) \supset H_n^m x]$$

	h	$\Delta(h)$	$\Delta^2(h)$	$\Delta^3(h)$	$\Delta^4(h)$	$\Delta^5(h)$	...
$H_3 0$	**f**	**t**	**t**	**t**	**f**	**t**	
$H_3 1$	**f**	**f**	**t**	**t**	**f**	**f**	
$H_3 2$	**f**	**f**	**f**	**t**	**f**	**f**	

Table 3: Pattern of truth values across revisions

In all models in which either '<' is not interpreted as a discrete linear order with a least element or there are not at least n elements, the revision sequence for H_n^m will settle at the empty set for the extension of H_n^m. Let us suppose that M is a model in which $LINORD(n)$ is satisfied. Then there are n objects. We may as well call them 0,1,2,..., $n-1$. As an example, the revision sequence for $H_3 x$ is as follows.[15] I will assume that the initial extension of H_3 is empty. After one revision, the extension of H_3 resets either to the empty set or to an initial <-segment containing no more than 3 objects. After at most 3 revisions, it resets to the empty set, and then falls into the pattern illustrated in table 3, until it reaches 3 elements, at which point it resets to empty after the next revision. With this observation, we can prove the following.

Lemma 63. *Each finite subset $Y \subseteq \mathscr{H}$ is a finite definition.*

For any H_n^m, we can construct a table of the pattern of truth values it takes over the n elements it applies to. This has an eventual period of $n+1$. For $0 \leq k < n+1$, let $Col(H_n^m)_k$ be a sentence that is true whenever the kth column of the table of patterns of truth values matches the current stage of revision. For H_n^m, there are $n+1$ columns and $Col(H_n^m)_k$ is defined as

$$\pm_0 H_n^m(0) \,\&\, \pm_1 H_n^m(1) \,\&\, \ldots \,\&\, \pm_n H_n^m(n),$$

where $\pm_j$ is nothing if $j < k$ and $\pm_j$ is '$\sim$' otherwise. For the example above, $Col(G_3)_0$ is

$$\sim H_3(0) \,\&\, \sim H_3(1) \,\&\, \sim H_3(2),$$

[15] I will drop the superscript and focus on the named 3 elements

and $Col(H_3)_2$ is

$$H_3(0) \,\&\, H_3(1) \,\&\, {\sim}H_3(2).$$

It is clear from the definition that if $Col(H_n^m)_k$ is true at a particular $M + h$, where h is reflexive, then $Col(H_n^m)_j$, where $j \neq k$, will not be true at $M + h$. Combinations of the $Col(H_n^m)$ sentences are what will interpret the boxed atoms for the countermodel. They will be used to obtain the desired pattern of truth values.

If A contains m distinct atoms and A has modal depth n, then the definition $\mathscr{D}$ will be the subset of $\mathscr{H}$ containing the definitions for $H_n^0, \ldots, H_n^{m-1}$. To define * we need some more notation. Let

$$[p_i] = \{(n+1) - k : k \leq n+1 \,\&\, M, w_k \Vdash p_i\}.$$

Let

$$(p_i)^* = \bigvee_{k \in [p_i]} Col(H_n^i)_k,$$

for each atom p_i in A and for all other atoms, $(q)^* = \bot$. If $[p_i] = \emptyset$, then $\bigvee_{k \in [p_i']} Col(H_n^i)_k = \bot$.

As before, we construct a sequence of hypotheses, $\langle h_i \rangle_{i \in \omega}$.

- $h_0 = \emptyset$
- $h_{i+1} = \Delta(h_i)$

After $n+1$ revisions, the H_n^ms will cycle back to empty extensions. If the H_n^m are empty at stage k, then at stage $k+n$, $M, h_{k+n} \models A_{bnf}^*$ iff $M, w_0 \Vdash A_{bnf}$. This is because h_{k+n} agrees with w_0 on the evaluation of all boxed atoms. For each j, there are k and r such that $h_{j+k} \equiv_{\mathscr{D}} h_{(r\cdot(n+1))+n}$, and $M, h_{j+k} \not\models_0^{\mathscr{D}} A_{bnf}^*$. Therefore, $M \not\models_0^{\mathscr{D}} A_{bnf}^*$, as desired.

5.2.2 First-order Solovay-type theorem

We can obtain a first-order version of the theorem using a similar technique. We must slightly alter the definition of a $\mathscr{D}$-interpretation. Instead of a propositional modal language, we start with first-order modal language. The atomic clause in the definition of a $\mathscr{D}$-interpretation should be as follows.

- $F(x_1, \ldots, x_n)^* = B(x_1, \ldots, x_n)$, where B is a formula in the language $\mathscr{L}^+_{\mathscr{D}}$ such that $x_1, \ldots, x_n$ are all and only the free variables in B.

The other clauses of the definition of $\mathscr{D}$-interpretation remain the same.

The theorem we wish to prove is the following, where our modal language contains no names.

Theorem 26. *RTQ* $\vdash A$ *iff* $\forall$ *simple finite* $\mathscr{D}$ $\forall \mathscr{D}$*-interpretations* $^* \models^{\mathscr{D}}_{\#} A^*$

As before, the soundness direction, the left-to-right direction, is immediate. The converse direction will take some more work.

We note the following.

Theorem 27 (*RTQ* completeness)**.** *RTQ is sound and complete with respect to constant domain Kripke frames in which every world has exactly one successor.*

Proof. This is proved using the methods of Hughes and Cresswell (1996), together with the fact that for every world w in the canonical model, w contains exactly one of $\Box A$ and $\Box{\sim}A$. □

We will prove the contrapositive, so we will assume that A is a sentence such that $RTQ \not\vdash A$. It follows that $RTQ \not\vdash A'$, where A' is the box normal form from A. By the completeness of RTQ, there is then a model M and a world w_0 such that $M, w_0 \not\models A'$. We will assume for notational simplicity that all of the predicates F_i have the same arity.

The model M assigns a pattern of extensions to the predicates that falsify A' at w_0. Since M is an RT model, each world in M has unique R successor. Let $\alpha = md(A) + 1$ and let $\beta = md(A)$. We need only look at the sequence of α worlds, starting from w_0. The extensions in those worlds can be represented as follows. The predicate F_j is assigned the extension X^j_n in w_n, for each $n \leq \alpha$. We will use this pattern of extensions to define the refuting hypotheses.

We will modify the set $\mathscr{G}$ from the propositional case to obtain the desired set of definitions. Abusing notation slightly, let $\mathscr{G}$ be the set of definitions, $G^m_n(\overline{x}) =_{Df} \Box^n G^m_n(\overline{x})$, for each n and m. As before, the extensions assigned to G^m_n repeat in a sequence over $n+1$-many stages. Each finite subset of $\mathscr{G}$ is itself a finite definition.

	w_0	w_1	w_2	...	w_β
F_1	X_0^1	X_1^1	X_2^1		X_β^1
F_2	X_0^2	X_1^2	X_2^2		X_β^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
F_m	X_0^m	X_1^m	X_2^m		X_β^m

Table 4: Pattern of extensions for predicates across worlds

As in the propositional case, it may be helpful to have an example of a cycle of extensions for a definition from $\mathscr{G}$. Take $G_3x =_{Df} \Box^3 G_3x$ and some model M. The definition cycles in the following pattern, presented in terms of extensions rather than pairs from hypotheses. We will drop the subscript on the predicate and corresponding superscript on the extension.

	h	$\Delta(h)$	$\Delta^2(h)$	$\Delta^3(h)$	$\Delta^4(h)$	$\Delta^5(h)$	...
Gx	X_0	X_3	X_2	X_1	X_0	X_3	
$\Box Gx$	X_1	X_0	X_3	X_2	X_1	X_0	
$\Box^2 Gx$	X_2	X_1	X_0	X_3	X_2	X_1	
$\Box^3 Gx$	X_3	X_2	X_1	X_0	X_3	X_2	

Table 5: Pattern of extensions assigned by hypotheses

For definition G_n^m, the initial hypothesis will return to its values over $sub(\mathscr{D})$ every $n+1$ revisions. The following chart lists the set of elements satisfying the formulas, $Gx, \Box Gx, \Box^2 Gx$, and $\Box^3 Gx$, at different stages of revision.

We now define the desired set $\mathscr{D}$ of definitions. We let $\mathscr{D}$ be the set of definitions for G_β^1, ..., and G_β^m. The desired $\mathscr{D}$-interpretation *is the one that assigns to each F_i the predicate G_β^i, with the appropriate variables. For the model, we take the domain of D and assign all predicates not occurring in $(A')^*$ an empty extension. Call this model M'. Define a sequence of hypotheses $\langle h_n \rangle_{n \in \omega}$ as follows.

$\models$	h	$\Delta(h)$	$\Delta^2(h)$	$\Delta^3(h)$	$\Delta^4(h)$	$\Delta^5(h)$	...
Gx	X_3	X_2	X_1	X_0	X_3	X_2	
$\Box Gx$	X_0	X_3	X_2	X_1	X_0	X_3	
$\Box^2 Gx$	X_1	X_0	X_3	X_2	X_1	X_0	
$\Box^3 Gx$	X_2	X_1	X_0	X_3	X_2	X_1	

Table 6: Pattern of sets satisfied by hypotheses

- $\langle \Box^k G^j_\beta \overline{x}, v\rangle \in h_0$ iff $v(\overline{x}) \in X^j_k$, where X^j_k is the extension of F_j in w_k.

- $h_{n+1} = \Delta_{M',\mathscr{D}}(h_n)$.

The sequence $\langle h_n\rangle_{n\in\omega}$ is a revision sequence. The desired hypotheses to falsify $(A')^*$ are all hypotheses that agree with M in the following sense. For each j, k such that $1 \leq j \leq m$ and $k \leq \beta$, X^j_k is the set of tuples satisfying $\Box^k G^j_\beta \overline{x}$ in $M' + h$ just in case M assigns X^j_k to F_j in w_k. Suppose h_n is such a hypothesis. For all p, $h_n \equiv_{\mathscr{D}} h_{p\cdot(\alpha)+n}$.

It remains to see that $\not\models^{\mathscr{D}}_0 (A')^*$. For this to be true, we need a model N and for each n, a hypothesis h such that $N, \Delta^n(h) \not\models (A')^*$. For the model we take M'. For each n, we want to pick a hypothesis that will yield one of the falsifying hypotheses after n revisions.

Now we can finally show that $M', \Delta^n(h) \not\models (A')^*$. The falsifying hypothesis $\Delta^n(h)$ agrees with the model M in the following sense: for each $k \leq \beta$, $M', v, \Delta^n(h) \models \Box^k G^p_\beta(\overline{x})$ iff $M, v, w_k \Vdash F_p(\overline{x})$, which holds just in case $M, v, w_0 \Vdash \Box^k F_p(\overline{x})$. Since $M, w_0 \not\models A'$, $M', \Delta^n(h) \not\models (A')^*$. It follows that $\not\models^{\mathscr{D}}_0 (A')^*$, so $\not\models^{\mathscr{D}}_0 (A)^*$. Since $\mathscr{D}$ is a simple finite definition, $\not\models^{\mathscr{D}}_{\#} (A)^*$, as desired.

We do not see how to replicate the proof of the first-order theorem for a definition that lacks boxes. In the propositional case, the box-free definition that we gave cycled through truth values, in other words, extensions with respect to a single element. We do not see both how to replicate these cycles with arbitrary extensions and how to ensure that the resulting definition is finite.

5.2.3 Variations of Solovay-type theorems

We can prove some variations on the Solovay-type theorem. We will start with a refinement of the propositional version.

We start with a language with infinitely many propositional variables and $\bot$. Partition this language into two sets, $\mathcal{A}$ and $\mathcal{B}$, with $\bot \in \mathcal{A}$. For all $p \in \mathcal{A}$, add the axiom $p \equiv \Box p$ to RT. We also strengthen the Nec rule to apply to all theorems, including those in whose proofs the new axioms appear. Call the resulting system $RT_{\mathcal{A}}$.

Given a Kripke model $M(= (W, R, V))$, partition W using the reflexive, symmetric, transitive closure of R. The cells of this partition are the R-blocks of M. Say that an RT model M is p-constant iff

$$\forall w, w'(M, w \Vdash p \ \Rightarrow \ (wRw' \vee w'Rw \ \Rightarrow \ M, w' \Vdash p)).$$

An RT model M is $\mathcal{A}$-constant if it is p-constant for all $p \in \mathcal{A}$.

Given a propositional language and a partition as above, we have the following.

Theorem 28. *$RT_{\mathcal{A}}$ is sound and complete with respect to the class of $\mathcal{A}$-constant RT models.*

For the Solovay-type theorem, we need to adjust the definition of a * interpretation. The atomic clause must be split into two cases.

- For $p \in \mathcal{A}$, $(p)^* \in Sentence_{\mathscr{L}}$
- For $q \in \mathcal{B}$, $(q)^* \in Sentence_{\mathscr{L}^+_{\mathscr{D}}}$

The * interpretations map atoms in $\mathcal{A}$ to base language sentences. Atoms in $\mathcal{B}$ can be mapped to sentences of the language expanded with $\Box$ and $\mathscr{D}$.

The theorem we prove is the following.

Theorem 29. *1. $RT \vdash A$ iff $\forall \mathscr{D}$ $\forall$ $\mathscr{D}$-interpretations* $\forall k \vdash^{\mathscr{D}}_0 (A^*)^k$*

2. $RT \vdash A$ iff $\forall$ simple finite $\mathscr{D}$ $\forall$ $\mathscr{D}$-interpretations * $\models^{\mathscr{D}}_{\#} (A^*)$

The proof of this proceeds along the same lines as the previous proof for the propositional case. The main difference is that for $p \in \mathcal{A}$, $(\Box^k p)^*$ can be replaced with $(p)^*$, which can be set to $\top$ or $\bot$, according to the refuting model. The set $\mathscr{D}$ of definitions is constructed using the subset of atoms from A in $\mathcal{B}$, rather than all the atoms in A.

For the first-order case, we can prove the theorem for $RTQ^{=}$ rather than just RTQ. We can also prove it for $RTQ^{=}$ with constants in the language, provided we adjust the equality axioms to cover constants. The definition of a $\mathscr{D}$-interpretation needs an additional clause if equality is in the language.

- $(s = t)^* = (s = t)$.

There is a final variation worth mentioning. Suppose we have a first-order language $\mathscr{L}$ with constants, equality, and possibly infinitely many atomic predicates. Let the set of the predicates in $\mathscr{L}$, including equality, be $\mathcal{A}$. Suppose we enrich $\mathscr{L}$ with a finite set of predicates $\mathcal{B}$ which is disjoint from $\mathcal{A}$. Call the new language $\mathscr{L}^+$. For each$n$-ary $G \in \mathcal{B}$, we associate the formula $G\overline{x}$, with n distinct variables, with a formula of $\mathscr{L}^+$, A_G, that both has exactly the n variables $\overline{x}$ free and may contain G. Let the set of such pairs be $\mathcal{D}$. We add to $RTQ^{=}$, the axioms $\forall\overline{x}(F\overline{x} \equiv \Box F\overline{x})$, for $F \in \mathcal{A}$ and the axioms $\forall\overline{x}(G\overline{x} \equiv \Box A_G(\overline{x}))$, for $(G(\overline{x}), A_G) \in \mathcal{D}$. We also strengthen the Nec rule to apply to theorems whose proofs include the new axioms. Call the resulting system $RTQ^{=}_{\mathcal{A},\mathcal{D}}$.

Say that an $RTQ^{=}$ model M is $(G(\overline{x}), A_G)$-compliant iff $\forall\overline{x}(G\overline{x} \equiv \Box A_G(\overline{x}))$ is valid in M. Say that M is $\mathcal{D}$-compliant if it is $(G(\overline{x}), A_G)$-compliant, for each $(G(\overline{x}), A_G) \in \mathcal{D}$. Say that an $RTQ^{=}$ model M is F-constant just in case, for all R-blocks of M, all worlds in a block assign the same extension to F. The definition for $\mathcal{A}$-constant models is similar.

There is then the question of whether, given such a language $\mathscr{L}^+$ and sets $\mathcal{A}$ and $\mathcal{D}$, the following holds.

Proposition 2. *$RTQ^{=}_{\mathcal{A},\mathcal{D}}$ is sound and complete with respect to the class of $\mathcal{D}$-compliant, $\mathcal{A}$-constant RTQ models with constant domains.*

We have not been able to prove this yet. Unlike the other cases, however, completeness is not needed for the proof of a Solovay-type theorem for $RTQ^{=}_{\mathcal{A},\mathcal{D}}$.

Before proceeding to the Solovay-type theorem, we need some terminology. Given a set $\mathscr{D}$ of definitions and a set $\mathcal{D}$ of pairs of n-ary predicate letters and formulas with n free variables, let us say that $\mathscr{D}[\mathcal{D}]$ agrees with $\mathcal{D}[\mathscr{D}]$ iff $G\overline{x} =_{Df} A_G(\overline{x})$ is in $\mathscr{D}$ just in case $(G(\overline{x}), A_G) \in \mathcal{D}$. The natural Solovay-type theorem is the following.

Proposition 3. *Let $\mathscr{D}$ be a simple finite definition and agree with $\mathcal{D}$. Then, for all sentences*

A of $\mathscr{L}^+$,

$$RTQ^{=}_{\mathcal{A},\mathcal{D}} \vdash A \Leftrightarrow \models^{\mathscr{D}}_{\#} A$$

Proof. Since $\mathscr{D}$ is a simple finite definition, $C^{\Box}_0$ is complete with respect to $S^{\#}$ validity. Therefore, the claim is equivalent to

$$RTQ^{=}_{\mathcal{A},\mathcal{D}} \vdash A \Leftrightarrow \vdash^{\mathscr{D}}_0 A.$$

For the left-to-right direction, we have already shown that C^{-} proves all axioms of $RTQ^{=}$. It remains to show that $C^{\Box}_0$ proves the additional axioms of $RTQ^{=}_{\mathcal{A},\mathcal{D}}$, namely the axioms $\forall\overline{x}(F\overline{x} \equiv \Box F\overline{x})$, for $F \in \mathcal{A}$ and the biconditionals for the pairs in $\mathcal{D}$. The former are consequences of the index shift rules. The latter follow immediately from the definition rules together with the box rules.

For the right-to-left direction, we have already shown that $RTQ^{=}$ proves all the theorems of C^{-}. We need to show that the steps of the earlier quasi-proof transformations can be carried out by $RTQ^{=}_{\mathcal{A},\mathcal{D}}$ if $C^{\Box}_0$ is used instead of C^{-}. The remaining rules are the index shift and definition rules.

Case: IS. This case is taken care of using the axioms $\forall\overline{x}(F\overline{x} \equiv \Box F\overline{x})$.

Case: DefI. This case is taken care of using the appropriate instances of the axioms $\forall\overline{x}(G\overline{x} \equiv \Box A_G\overline{x})$.

Case: DefE. This is similar to the DefI case. We conclude that, given the assumptions, if $\vdash^{\mathscr{D}}_0 A$, then $RTQ^{=}_{\mathcal{A},\mathcal{D}} \vdash A$, which completes the proof.

□

We conclude with one final variation for the *RT* proof. In that proof, the size of $\mathscr{D}$ depended on the number of atoms in A. If we permit the first-order language to contain infinitely many names, then for each A such that $RT \nvdash A$, we can specify a finite definition $\mathscr{D}$ that contains exactly one definition such that $\not\models^{\mathscr{D}}_0 (A)^*$. For each n and m, let G^m_n be defined as follows, where $Distinct(n)$ is a sentence saying that $a_i \neq a_j$, if $i \neq j$.

$$G^m_n(x) =_{Df} Distinct(n) \,\&\, ((x = a_1 \,\&\, \Box^m G^m_n(a_1)) \vee \ldots \vee (x = a_n \,\&\, \Box^m G^m_n(a_n)))$$

Each singleton containing a G_n^m is a finite definition. If A has n atoms and modal depth m, then each atom p_i is interpreted as $G_n^m(a_i)$. Let the model M have a countable domain in which $Val_M(a_i) \neq Val_M(a_j)$, if $i \neq j$. As in the earlier proof, the distribution of truth values for the atom p_i in the falsifying RT model specifies the appropriate hypothesis for $G_n^m(a_i)$. The proof then proceeds in much the same way as before.

6.0 FIELD'S THEORY

Hartry Field has recently proposed a theory of truth that builds on the strong Kleene fixed-point theory by adding a new conditional, $\rightarrow$, to the logic.[1][2] In (§6.1), I present some of the criticisms of the strong Kleene theory of truth that demonstrate some of the problems with the conditional of strong Kleene logic. Following this, I will present some of the details of Field's conditional (§6.2), covering both its technical details and philosophical motivations. With that background in place, I will argue that Field's conditional is inadequate, both by my criteria and by his own (§6.3). Part of the problem with Field's conditional is that it is neither clear what the logic of the conditional is nor whether the conditional is adequate for its roles. I will close by presenting a framework for investigating the logic of the conditional (§6.4).

6.1 BACKGROUND

Field's approach builds on the strong Kleene fixed-point approach, which Field thinks is on the right track with respect to truth. It does this by adding a new conditional to the logic. The fixed-point approach to truth has some notable features. It has a so-called *transparent truth predicate*. Its truth predicate is such that for all sentences A, A and $T(\ulcorner A \urcorner)$ have the same semantic value:

$$||A|| = || T(\ulcorner A \urcorner) ||.$$

[1] I will concentrate on Field's view as it is presented in Field (2008).

[2] In this chapter, I will use '$\rightarrow$' and '$\leftrightarrow$' for Field's conditional and biconditional, respectively. I will continue using '$\supset$' and '$\equiv$' for material conditional and biconditional, respectively.

This extends to all extensional contexts C.

$$||C(A)|| = ||C(T(\ulcorner A \urcorner))||.$$

This latter principle has become known as the *Intersubstitutivity Principle*, or (IP). Additionally, the construction of the fixed-point models is straightforward and can be adapted to work with any monotonic scheme.[3] Some classes of fixed-points have elegant, complete proof systems.[4]

The fixed-point approach to truth has some notable defects. As pointed out in Kripke (1975), the liar sentence is not in the extension of truth nor in its anti-extension, but there is no way to say truly, in the object language, that the liar is not true. Gupta (1984) presents several compelling objections to the fixed-point view of truth. There are four on which I wish to concentrate, for reasons that will become clear shortly. Gupta's objections center on how the fixed-point approach delivers incorrect results.

The first objection is that the fixed-point models never make all the T-sentences true, and some T-sentences are never true. For example, the liar's T-sentence,

$$Ta \equiv \sim Ta,$$

cannot be in any fixed-point. The T-sentence for a liar is never assigned **t**.

The second and third objections go together as both concern the law of non-contradiction. The second objection is that the law of non-contradiction,

$$\forall x \sim (Tx \,\&\, \sim Tx),$$

is paradoxical when a liar sentence, or another instance of vicious self-reference, is present in the language. A sentence is *paradoxical* iff there is no fixed-point in which it is true. The liar is paradoxical, because it must always take the value **n** and so can never be in a fixed-point. The law of non-contradiction will take the minimum value of its instances, and the instance with the liar,

$$\sim (Ta \,\&\, \sim Ta),$$

[3] See Kripke (1975) or Gupta and Belnap (1993) for details.
[4] See Kremer (1988) for details.

will receive the value **n**. The quantified form of the law of non-contradiction will then receive the value **n** as well.

The third objection is that even if there is no vicious self-reference in the language, the law of non-contradiction is still pathological. A sentence is *pathological* iff it is not in the minimal fixed-point. To set this up, let L be a first-order language with some predicates interpreted classically by the ground model and expand it to a language L^+ by adding a truth predicate and quotation names for all and only the sentences of L^+. The sentences of L^+ and the quotation names to be added are defined by simultaneous induction. Let us call L^+ the *minimal syntactic expansion of* L. There are no other syntactic resources in the language besides the quotation names. This language lacks vicious self-reference because the only quotation names available are those for sentences that are constructed via the standard inductive definition of wffs. The law of non-contradiction has itself as an instance, so it cannot get into the fixed-point until it is evaluated as **t**. This cannot happen in the minimal fixed-point, so the law of non-contradiction is pathological.

The fourth objection is that the fixed-point approach, particularly the minimal fixed-point, is wrong about logical reasoning with truth. To see this, we consider the following set of sentences.

A₁ $T(\ulcorner B \urcorner)$

A₂ $\sim T(\ulcorner B \urcorner)$

B $(T(\ulcorner A_1 \urcorner) \vee T(\ulcorner A_2 \urcorner)) \,\&\, \sim(T(\ulcorner A_1 \urcorner) \,\&\, T(\ulcorner A_2 \urcorner))$

Sentence B says that one of A_1 and A_2 is true, but not both, while A_1 and A_2 offer contradictory assessments of the status of B. A_1 and A_2 are contradictory, so one expects exactly one of them to be correct, which means that B is true, hence A_1 is true and A_2 is false. The minimal fixed-point says that all of the sentences get the value **n**. Some classes of fixed-points do better than the class of minimal fixed-points on delivering intuitively correct verdicts about consequence. The class of greatest intrinsic fixed-points, for example, get this particular example correct.

There are two other features of the minimal fixed-point theory of truth worth pointing out. First, it delivers incorrect verdicts on logical consequence. If a is a liar and b is a truth

teller, then the minimal fixed-point theory says,

$$Tb \models Ta,$$

because liars and truth tellers are always evaluated as **n** in the minimal fixed-point.[5] This entailment verdict seems to be incorrect. Another example of an incorrect verdict is that the minimal fixed-point theory says that the law of non-contradiction and the liar are equivalent.[6] Both receive the value **n** in all minimal fixed-points. Second, the consequence relation of the class of minimal fixed-points is not axiomatizable.[7] I will return to both points in §6.2.1

Visser (2004, 204-205) summarizes Gupta's criticisms of the fixed-point theory as follows. Visser says that Gupta's criticisms are misplaced because they focus on the wrong kind of negation. The only negation available to the then current fixed-point approaches was choice negation. The criticism concerns a stronger negation, one that could be defined in terms of a conditional, such as $\rightarrow\bot$. The point about T-sentences also indicates a weakness in the fixed-point theory's conditional. Visser says, "If this is true, Gupta's whole criticism comes to: [fixed-point theories] lack a good conditional."[8]

The fixed-point theories based on weak or strong Kleene logic have weak conditionals. For example, there are no valid conditionals using the strong Kleene material conditional. The material conditional $A \supset A$ is equivalent to $\sim A \vee A$, which receives the value **n** whenever A does. The conditional does not obey the substitution of equivalents,

$$A \equiv B \models C(A) \equiv C(B),$$

where C is any extensional context. Let the context be '$D\,\&\,$' and suppose that A and B both receive the value **t** while D receives **n**. The premises are satisfied, but the conclusion is not.

Field is a proponent of the fixed-point approach to truth, in particular the minimal strong Kleene fixed-point approach, because it provides a transparent truth predicate. He shares the view that the material conditional of strong Kleene logic is too weak. The primary

[5] Gupta and Belnap (1993, 99) attributes the discovery of this to Kremer (1986) and Visser (1989).
[6] This was pointed out to me by Anil Gupta.
[7] See Kremer (1988) for the proof.
[8] Visser (2004, 205)

failings he sees with it are that it does not obey the substitution of equivalents and that it does not guarantee the validity of the T-sentences. Field thinks that the way to fix the faults of the fixed-point approach are to augment strong Kleene logic with a new conditional.[9]

Field (2008) introduces a new conditional to remedy some of the deficiencies of the strong Kleene fixed point approach to truth. He defines this conditional via a revision procedure.[10]

6.2 OVERVIEW OF FIELD'S CONDITIONAL

Field takes the basic strong Kleene fixed-point theory of truth and augments it with a new conditional, $\rightarrow$, which he thinks fixes the problems of the basic theory. Field's new conditional is defined via a revision construction, the details of which I will now sketch.[11] We fix an interpretation for the language without the truth predicate and conditional. The revision proceeds in a two-step way. Given an interpretation of conditional sentences, the least fixed-point for truth is constructed, treating all conditionals as atoms for the duration of the fixed-point construction. The values of sentences in the fixed-point are used to revise the interpretation of the arrow according to the following rules for successor and limit stages.

- $v_0(A \rightarrow B) = \mathbf{n}$
- $v_{\alpha+1}(A \rightarrow B) = \begin{cases} \mathbf{t} & v_\alpha(A) \leq v_\alpha(B) \\ \mathbf{f} & v_\alpha(A) > v_\alpha(B) \end{cases}$
- $v_\lambda(A \rightarrow B) = \begin{cases} \mathbf{t} & \exists\alpha < \lambda \forall\beta(\alpha \leq \beta < \lambda \Rightarrow v_\beta(A) \leq v_\beta(B) \\ \mathbf{f} & \exists\alpha < \lambda \forall\beta(\alpha \leq \beta < \lambda \Rightarrow v_\beta(A) > v_\beta(B) \\ \mathbf{n} & \text{otherwise} \end{cases}$

The initial interpretation of all conditionals is **n**. At successor stages, a conditional is revised to **t** if the value of its antecedent at the previous stage is no greater than the value of its

[9] There is no indication in Field (2008) that Field read Visser (2004) or that he is concerned to respond to Gupta's objections. Nonetheless, I think that presenting Gupta's objections and Visser's response sets up one way of examining Field's proposal. I will return to this point in §6.3.1.

[10] Field introduced precursor ideas in articles, such as Field (2004) and Field (2003). I will focus on the presentation in Field (2008), since it contains the most sustained discussion of the philosophical issues and Field seems to endorse the theory presented there.

[11] I will omit some details needed to account for the quantifiers. The points I wish to make about the conditional can be made without involving the details of quantification.

consequent at the previous stage, and otherwise the conditional is revised to **f**. Limit stages use the rule that conditionals that have stabilized go to their stable values while unstable ones are set to **n**.

Note that the revision of the conditional must be treated separately from the construction of the fixed-point for truth. This is because the construction of the fixed-point may fail to reach a fixed-point if the logical operators lack a certain monotonicity property, a property which the successor stage truth table for Field's conditional lacks.

The revision process for Field's conditional eventually enters a loop of repeating interpretations for the conditional. There are stages in the loop that serve as particularly nice interpretations for the conditional. I will call these *reflection stages*, echoing the terminology of Gupta and Belnap (1993).[12] Reflection stages are nice in the sense that they ensure that conditionals do not upset the logic of the other connectives.[13] The fixed-points above reflection stages are Field's preferred models for his theory of truth.

Field defines the semantic concept of *ultimate value* as follows. The ultimate value of of an atom in a model M is the value, $\mathbf{t}, \mathbf{n}$, or $\mathbf{f}$ that the atom takes in the least fixed-point over a reflection stage. The ultimate value of a conditional $A \rightarrow B$ in a model M is defined in the same way as for atoms. The ultimate value of $\sim A$ is $\mathbf{f}$ if the ultimate value of A is $\mathbf{t}$, $\mathbf{t}$ if the ultimate value of A is $\mathbf{f}$, and $\mathbf{n}$ otherwise. The ultimate values of $A \vee B$ and $A \,\&\, B$ are defined as the maximum and minimum, respectively, of the ultimate values of A and B.[14] Consequence can be defined as preservation of ultimate value $\mathbf{t}$.[15]

Earlier, I pointed out three defects of strong Kleene logic, two which centered on the material conditional and one on the liar. Field's conditional remedies all three defects. Field's biconditional is defined as

$$A \leftrightarrow B =_{Df} (A \rightarrow B) \,\&\, (B \rightarrow A).$$

Field's models validate all of the T-sentences that use his biconditional as well as the rule form of substitution of equivalents. To fix the problem regarding the status of the liar, Field

[12] Field calls the indices of these stages "acceptable ordinals."
[13] Field proves this as a theorem he calls "the fundamental theorem." See Field (2008, 257-258) for a proof.
[14] The quantifiers can be treated as suprema and infima.
[15] I will delay discussion of the laws and inferences validated by this semantics until §6.4.

defines an operator D, read as "determinately," in terms of his logical resources:

$$DA =_{Df} A \,\&\, {\sim}(A \rightarrow {\sim}A).$$

Informally, DA says that A has value $\mathbf{t}$ at the current and previous stages. Liars, such as Ta, where $a = \ulcorner{\sim}Ta\urcorner$, turn out to be not determinately true, or ${\sim}DTa$ is true. New liar sentences can be formed using the D operator, such as Td, where $d = \ulcorner{\sim}DTd\urcorner$. Td turns out to be not determinately determinately true. More pathological sentences can be formed and evaluated using iterations of the D operator. This point is an important one to which I will return, with examples, in §6.3.2.

In addition to remedying the three defects of the strong Kleene theory that I pointed out, Field's theory of truth maintains one of the strong Kleene theory's notable features. The truth predicate of Field's theory obeys the Intersubstitutivity Principle.

Field's canonical models are defined with respect to single revision sequences in the sense that, given an interpretation of the initial language, only one sequence is used to define Field's canonical model for that language. Additionally, only one limit rule is used, a constant limit rule that is particularly simple. I will come back to both of these points in §6.3.2.

This is sufficient background on the formal aspects of Field's theory of truth. I will now proceed to my main criticisms of Field's theory.

6.2.1 General logic

Field thinks that we should adopt his logic, strong Kleene with his new conditional, as our general, all-purpose logic. He says, "the idea is that we can take the paracomplete logic to be our single all-purpose logic."[16] Within certain domains, we can assume excluded middle, and so Field's logic will collapse to classical logic in those domains.

Field's suggestion is that we should revise our logical and inferential behavior to accord with what his logic says is valid. Field says the following.

[16] Field (2008, 15)

> The situation with truth is similar [to that of switching from Euclidean to Riemannian geometry]: here the "old theory", involving both classical logic and the naive theory of truth, is simply inconsistent. Indeed it's trivial: it implies everything, e.g. that the Earth is flat. If you don't want to be committed to the view that the Earth is flat you need a theory that differs from the naive theory in basic principles, either principles about truth or principles about logical matters more narrowly conceived. If giving up those basic principles involves a "change of meaning", so be it: for then the "old meanings" aren't really coherent, and they need changing.[17]

Field thinks that the naive theory of truth, the set of T-sentences, is our background view of truth. In combination with classical logic it results in an inconsistent theory. The logical and inferential behavior governed by that theory needs to be changed, and Field thinks that his logic is the one to adopt.

The problem I want to raise for Field's proposal is that it is not clear what the suggested change is. The only characterization Field provides of his logic is via the construction of a canonical model of his theory of truth. The truths of the logic are whatever is true in the minimal fixed-points above the reflection ordinals of the sequences. There is no complete axiomatization of even the propositional fragment. Field's preferred consequence relation is defined via substitution of formulas in the language of set theory and uses the canonical model construction. Welch (2008) shows that the resulting consequence relation is highly complex.

Models of set theory are complex, so, one might think that starting with a simpler ground model and carrying out the construction would lead to something more tractable. Peano arithmetic serves as a standard background syntactic theory for work on truth, so the standard model of arithmetic is a possible starting model. This, however, will not work either. McGee (2010) showed that the set of validities of the conditional and truth with arithmetic as the background syntactic theory is also complex, Π^1_2, in fact. This rules out axiomatization of the conditional in the theories of truth built over arithmetic.

The extension of the consequence relation is complex for Field's canonical models of his theory of truth. Field has the problem of defining the set of arguments that are really valid as opposed to simply validated by his particular formal semantics. In response to this, he says the following.

[17] Field (2008, 16-17)

> In addition, the set of "logically valid inferences" will have an extremely high degree of non-computability.... It might be better to adopt the view that what is validated by a given version of the formal semantics outruns "real validity": that the genuine logical validities are some effectively generable subset of those inferences that preserve value 1 in the given semantics. If one adopts this viewpoint then there would doubtless be some arbitrariness in which effectively generable subset to choose, but that seems perfectly acceptable unless one wants to put high (and I think unreasonable) demands on the significance of the distinction between those inferences that are logically valid and those that aren't.[18]

There are at least two questions to which we, as logicians, want answers.

- What is the propositional logic of the strong Kleene scheme with the conditional?
- What is the quantificational logic of the strong Kleene scheme with the conditional?

The question of the propositional logic of the conditional is, perhaps, the more pressing of the two. I take it that Field agrees that these are pressing questions, since in a discussion of ways of generalizing continuum-valued semantics, he says, "My ultimate interest is less in the semantics than in the logic that the semantics validates...."[19] The unique behavior of Field's conditional is tied to the presence of pathological sentences in the language, which requires a truth predicate. It is not obvious what the proper way to uncover the propositional logic of the conditional is. In §6.4, I will propose a framework for investigating the propositional logic apart from the truth predicate.

I said that Field does not provide an axiomatization of his conditional. Field (2008, Ch. 17.4) lists some valid conditionals and valid arguments, but these provide, at best, a partial grip on how to use the conditional. The list covers many axioms and inference forms, but Field does not make any claim that it is complete.

Indeed, Field defines a logic for truth in terms of his theory of truth built over a standard model of arithmetic. This logic is defined in terms of substitutions of sentences of the language with arithmetic vocabulary and truth for schematic letters.[20] The resulting logic is not axiomatizable.[21] This incompleteness is one of two primary problems with the list of axioms and inferences Field provides. This logic is defined in terms of the theory of truth, which builds in extra complexity. In §6.4, I will define a framework in which to study the

[18] Field (2008, 277)
[19] Field (2008, 232).
[20] See Field (2003) for details on the definition of this logic.
[21] As mentioned earlier, see Welch (2008) and McGee (2010) for the proofs.

logic of the conditional without the truth predicate and syntactic theory.

Field does not provide an axiomatization of his logic. He does, however, present a semantic characterization of it. The devil's advocate will, at this point, ask what is wrong with a semantic characterization alone. Having both an axiomatization as well as a description of a class of models would be nice, but one will suffice for philosophical purposes as much as the other. Our devil's advocate says the lack of an axiomatization for the logic is a minor flaw, rather than a major failing.

Further, we have a semantic characterization of the standard arithmetic with no complete axiomatization possible.[22] In the realm of modal logics, there are natural classes of frames that do not have a complete axiomatization.[23] These properties are not glaring flaws for either the modal frames or the standard model of arithmetic. Why should a similar property be a problem for Field's logic?

If the only characterization of Field's logic is the canonical model construction, then the lack of an axiomatization is a problem. As McGee remarks, the construction is complicated and there is no guarantee of success in finding the models to decide whether an argument is valid or not.[24] I will provide an alternative semantic characterization of the propositional logic, which I think lessens, but does not eliminate, the criticism. A complete axiomatization of the propositional fragment would clear up what the logic is.

The two examples, the standard model and modal frames with no axiomatization, can both be used to clarify the objection. I will start with the modal logic example.

There are classes of frames that do not have complete axiomatizations. The lack of axiomatizations does not impugn their interest to modal logicians. It does, however, become a problem if one of these logicians suggests that the logic of one of these classes of frames should be adopted as a general guide to our reasoning. We can get a feel for some of the argument forms that are valid, but these forms do not exhaust the valid patterns. Each time a new pattern is employed, one would have to check for invalidity by constructing a countermodel with no guarantee of success. At this point, the lack of an axiomatization makes it hard to follow through on the logician's suggestion to adopt the logic as our general,

[22] Completeness is possible if we allow the ω-rule, but I set that aside here.
[23] See Blackburn et al. (2002, 213-218) for discussion and proofs.
[24] McGee (2010, 427)

all-purpose logic.

Now I will turn to the arithmetic example. The standard model of arithmetic has a salient difference with Field's logic. The standard model does not purport to serve as a general guide for our reasoning. It does serve as a guide to how we should compute with natural numbers. Logic is frequently viewed as basic or general in a way that arithmetic is not.

The issue is made more difficult by Field's view that the truth predicate is a logical notion. He says that the notion of truth used in real soundness proofs is "subject to rules which themselves have a logical status."[25] He goes on to say, "The question of the soundness of classical logic is really just part of the question of the soundness of the overall 'logic' employed in the theory, including the 'logic of truth'."[26] The logic of truth requires a background syntactic theory, which Field takes to be either a standard interpretation of Quine's protosyntax, concatenation theory, or the standard interpretation of Peano arithmetic. The interpretation of the syntactic theory must be fixed, so it is logical as well.[27]

I think that Field is too hasty in treating arithmetic as logical on the basis of the idea that the syntactic theory is logical. It is plausible to keep the interpretation of the syntactic theory fixed so that strangeness in the behavior of the truth predicate is not due solely to a non-standard interpretation of the syntactic vocabulary. It is too quick to require that arithmetic be the syntactic theory for the general logic of truth. There are weaker options, such as that of Kremer (1988), which uses quotation names for sentences as the basic syntactic theory for the logic of truth.[28] The interpretation of the quotation names is standard, with

$$I(\text{`}A\text{'}) = A,$$

for every sentence A of the language. The quotation names can be viewed as logical and their interpretation held fixed, while arithmetic vocabulary need not be.

This response provides a small wedge between asking about Field's logic of truth, which

[25] Field (2008, 191)
[26] Field (2008, 191)
[27] This is based on Field's comments on why he rejects the possibility of using an ω-inconsistent arithmetic theory as the syntactic theory.
[28] As far as I have been able to tell, Gupta (1982) is the first published use of quotation names for the syntactic theory, rather than arithmetic or concatenation theory, in the study of truth.

includes his conditional and quotation names, and Field's theory of truth, which is built on arithmetic or set theory. The arithmetic theory need not be viewed as logical, so we can maintain a difference between being unable to axiomatize the standard model of arithmetic and being unable to axiomatize the logic of truth. I conclude that the arithmetic example does not bolster the case for Field's logic.

Additionally, it should be possible to separate the logic of the conditional from the logic of truth. The revision sequences for the conditional are sensitive to the resources of the language, because the conditionals are sensitive to the revisions of values of their parts. In Field's model construction, the only atoms whose semantic values change are those that involve the truth predicate. This might give the impression that we cannot inquire about the logic of Field's conditional without including the truth predicate in the language. I will provide a framework showing that this is not the case. The propositional logic can be separated from the truth predicate.

The second problem with the list is that it does not provide enough detail on when we can introduce the conditional. In general, the rule of conditional proof is not valid. The following restricted form is valid.

$$\frac{\Gamma, A \vdash B}{\Gamma, A \vee \sim A \vdash A \rightarrow B}$$

This is helpful, but it does not exhaust the possibilities. There are many cases in which we may want to introduce a conditional when the antecedent of said conditional does not obey excluded middle. For example, in showing that the liar is not determinately true, we would, presumably, reason from the premiss that Ta to the intermediate conclusion that $\sim Ta$, and conclude that $Ta \rightarrow \sim Ta$. From this we infer that the liar is not determinately true, via the strong Kleene rule for disjunction and de Morgan's laws.

Additionally, the above form of conditional proof is valid for the strong Kleene material conditional, which can claim the same benefits as Field's conditional on this point. The virtues of Field's conditional lie in the ways in which it gives us more than the strong Kleene material conditional. It might be that, upon further investigation, there is nothing more general to say about conditional proof for Field's arrow, but this is not clearly the case. There are arguments that seem to use a form of conditional proof for which we do not have

excluded middle for the antecedent of the conditional. In the following, the premiss sequents are valid and the conclusion sequents are valid.

$$\frac{A \to B, B \to C \vdash A \to C}{A \to B \vdash B \to C \to .A \to C}$$

$$\frac{A \to B, B \to C \vdash A \to C}{B \to C \vdash A \to B \to .A \to C}$$

These are valid for arbitrary A, B and C.

Conditional proof may not be the correct way to account for the validity of these inferences. In that case, it would be useful to have some pattern to help us figure out when a conditional can be introduced.

All valid statements of strong Kleene consequence are valid conditionals, so there are some conditionals involving antecedents that may not satisfy excluded middle in the axiomatization. However, if the reasoning does not proceed along strong Kleene lines, because, for example, of the use of modus ponens on a conditional at some point, we lose the opportunity to conclude with a conditional introduced by conditional proof.

An example will help illustrate the difficulties here. Consider the following arguments.

$$(1)\ \frac{A, B, C \vdash A}{A \vdash B \,\&\, C \to A}$$

$$(2)\ \frac{A, B, C \vdash A}{A, B \vdash C \to A}$$

$$(3)\ \frac{A, B, C \vdash A}{\vdash A \,\&\, B \,\&\, C \to A}$$

$$(4)\ \frac{A, B, C \vdash A}{A \vdash B \to .C \to A}$$

$$(5)\ \frac{\dfrac{A, B, C \vdash A}{A \vdash B \to (C \to A)}}{\vdash A \to (B \to (C \to A))}$$

$$(6)\ \frac{\dfrac{A, B, C \vdash A}{A, B \vdash C \to A}}{B \vdash A \to (C \to A))}$$

Inferences (1) through (4) are all valid. The initial inferences in (5) and (6) are both valid, while the final inferences in each are invalid. All of the valid inferences in (1)-(6) above remain valid even when A, B, C, and D are pathological, and so do not obey excluded middle.

In some ways, it is a virtue of Field's conditional that it does not obey conditional proof unrestrictedly. It is essential to Field's way of responding to variants of Curry's paradox. In other ways, however, it is a vice. It makes using the conditional rather difficult.

Conditional proof is not unrestrictedly valid. This is related to two features of Field's consequence relation. The first is that the official consequence relation is based on **t**-preserving strong Kleene logic. The second is that the semantic clause for the truth of a conditional in successor stages corresponds to **tf**-preserving strong Kleene consequence.

This second point bears some additional explanation. A conditional $A \rightarrow B$ will be valid if B can be obtained from A using **tf**-preserving strong Kleene reasoning alone. This means that B is a consequence of A in virtue of strong Kleene structure, that is, in terms of $\&, \vee$, and $\sim$.

One cannot infer from a statement of **t**-preserving consequence that the consequence will be **tf**-preserving. It is, however, this difference that separates Field's official consequence relation from his conditional.

The primary area of application for Field's conditional is the theory of truth. Within this area, excluded middle will frequently be unavailable, if Field's own theory is correct. The restricted form of conditional proof will be of little use in the area to which Field's logic is most applicable.

I will now proceed to some of the negative features of the logic.

6.3 NEGATIVE FEATURES OF FIELD'S LOGIC

Let us bracket the question of what the exact logic of Field's conditional is, important though that question is. There is a further question of whether the logic delivered is one that solves the problems that we had with the fixed-point approach. For this discussion, I appeal only to those features that will be a part of whatever Field's logic turns out to be.

Field introduces his conditional to solve the problems of the strong Kleene fixed-point approach to truth. He has in view a particular subset of problems, and his conditional does solve these. However, if one was unsatisfied with the fixed-point view in general, then, as I

will argue, one will not be satisfied with Field's view of truth.

These criticisms concern not only what might properly called the "logic," namely the sentential connectives and quantifiers, but also the truth predicate. I think that I am licensed in using the truth predicate in my argument for the following reason. Field introduced the conditional primarily to improve the fixed-point theory of truth, and it is in the context of the truth predicate that we should ask whether the logic delivered solves the problems.

6.3.1 Responding to Gupta's criticisms

In §6.1 I provided four criticisms that Gupta leveled against the fixed-point approach to truth. Visser's response to these was that they would go away once a good conditional was added to strong Kleene logic and that conditional was used to define a negation. Let us examine whether that hypothesis bears out if we use Field's conditional.

The first criticism is successfully answered. All the T-sentences are valid when formulated with Field's conditional. This is not surprising, given that it was a guiding principle in the formulation of the conditional.

For the next three criticisms, define a new negation: $\neg =_{Df} \rightarrow\bot$. Note that the addition of $\rightarrow$ provides new resources for the logic, such as the defined negation, which is distinct from the old negation, $\sim$. The old negation, however, is still available, and problems stemming from it are unchanged.

Let us turn to the second criticism, that the law of non-contradiction is paradoxical when other paradoxical sentences are in the language. To evaluate this criticism, let us define the modified law of non-contradiction as follows.

Modified LNC $\forall x\neg(Tx \,\&\, \neg Tx)$

We need to see what the status of the modified LNC is under the different conditions. There are other sentences like the modified law of non-contradiction, such as the modified determinate law of non-contradiction.

Modified DLNC $\forall x\neg(DTx \,\&\, \neg DTx)$

In general, there will be one such form of the law of non-contradiction for each iteration of the D operator. These will be present even in the minimal syntactic expansion because D is

definable from the logical vocabulary including the conditional. There are then two cases to consider: one in which we add quotation names for the D^{n}LNC sentences and one in which we do not.

The modified law of non-contradiction says nothing is true at the previous stage stage and non-true two stages back. The difference with regular law of non-contradiction is due to the difference in the ways the connectives are evaluated in the course of the construction of the canonical models. The strong Kleene connectives are evaluated after the revision of the conditional. Suppose that A oscillates between **n** and **t** at successor stages. $\sim A$ will oscillate between **n** and **f** at successor stages, while $\neg A$ will be stably **f**. At limit stages, A and $\sim A$ will be set to **n** while $\neg A$ will be set to **f**. These will be the respective ultimate values of A, $\sim A$, and $\neg A$ as well.

Due to these features of $\neg$, I do not think that the prospects for the modified law of non-contradiction in the presence of vicious self-reference look good. To help make my case, I will define iterated Curry sentences c_n as follows. To cut down on parentheses, the grouping of the conditionals will be to the right. First, I need to define iterated arrow notation.

- $A \rightarrow_0 B =_{Df} B$
- $A \rightarrow_{n+1} B =_{Df} A \rightarrow .A \rightarrow_n B$

Definition 25 (Iterated Curry)**.** *For each $n \geq 1$, c_n names the following sentence.*

$$Tc_n \rightarrow_n \bot$$

The revision pattern of iterated Curry sentences is easy to figure out. After n+1 revisions, Tc_n falls into the following pattern.

$$\underbrace{\mathbf{t} \ldots \mathbf{t}}_{n} \mathbf{f}$$

After $n+1$ revisions, $Tc_n \rightarrow \bot$, which is $\neg Tc_n$, falls into a similar pattern.

$$\mathbf{t} \underbrace{\mathbf{f} \ldots \mathbf{f}}_{n}$$

After n+1 revisions, $Tc_n \,\&\, \neg Tc_n$ falls into the same pattern as $\neg Tc_n$ and after n+2 revisions, $\neg(Tc_n \,\&\, \neg Tc_n)$ falls into the same pattern as Tc_n.

In a language whose vicious self-reference is exhausted by the set of iterated Curry sentences, the modified law of non-contradiction will be false at most successor stages. This is because the Curry sentences oscillate in truth value across stages. The Curry sentences can have periods of arbitrary finite length, so the determinate modified laws of non-contradiction will be falsified as well. Different Curry sentences will falsify the various laws of non-contradiction at different stages, so at limit stages, the laws of non-contradiction and all of the iterated Curry sentences will be set to **n**, which will also be their ultimate values.

Contrary to what Visser's suggestion would have us believe, the modified law of non-contradiction is evaluated as **n**, which is the same as the basic strong Kleene fixed-point evaluation. For large stretches of the construction, however, the modified law of non-contradiction is evaluated as **f**.

For languages with as much vicious self-reference as the set of iterated Curry sentences, then the modified law of non-contradiction will not be stably **t**. There is no reason to think that the addition of further vicious self-reference will result in a better evaluation of the modified law of non-contradiction. I conclude that Field's logic does not respond to the second of Gupta's objections.[29]

The fourth objection is easier to evaluate since it lacks quantifiers. The objection said that the fixed-point approach evaluated this set of sentences incorrectly.

A₁ $T(\ulcorner B \urcorner)$

A₂ $\sim T(\ulcorner B \urcorner)$

B $(T(\ulcorner A_1 \urcorner) \vee T(\ulcorner A_2 \urcorner)) \,\&\, \sim(T(\ulcorner A_1 \urcorner) \,\&\, T(\ulcorner A_2 \urcorner))$

The fixed-point approach assigns them all **n**, when an evaluation of **t** for A_1, **f** for A_2, and **t** B makes better sense of naive reasoning here.

Let us replace '$\sim$' with '$\neg$' in A_1, A_2 and B. The resulting evaluation will cause all three sentences to oscillate together in the pattern **ft** after one revision. They will all receive ultimate value **n**. This is due, in large part, to the fact that the modified form of A_2 is "about" something different than A_1. At successor stages, A_2 evaluates the previous stage

[29] I will omit discussion of the third objection here because it is much tricker to work out the patterns for the various forms of the law of non-contradiction, and the basic point has been already made.

value of B, while A_1 evaluates the current stage value. At each successor stage, A_1 and A_2 are both right, but they are not in danger of saying conflicting things.

Using Field's conditional to implement Visser's suggestion gives inadequate results. Field's conditional successfully responds to one of the four objections. Its answers to the others are no better than those in the basic strong Kleene theory. It substantially increases the complexity in evaluating the objections, so its responses are, in that way, worse.

Visser's idea, implemented with Field's conditional, does not work as a response to Gupta's objections. Beyond this, it is not obvious what further conclusions to draw. There are three options.

1. Visser was wrong that Gupta's objections focus on the lack of an adequate conditional in strong Kleene logic.
2. Visser's idea of defining a better negation from an adequate conditional was a poor idea.
3. Field's conditional is not an adequate conditional.

I think that the first is not the proper conclusion to draw. Strong Kleene logic lacks an adequate conditional. Gupta's objections highlight a weakness in the strong Kleene negation, but there are standard connections between conditionals and negations. In many contexts, a conditional $\rightharpoonup$ can be used to define a negation '$-$' in the following way: $-=_{Df} \rightharpoonup \bot$, where $\bot$ is a propositional constant interpreted via $\mathbf{f}$. If '$\supset$' is the strong Kleene material conditional, then '$\supset \bot$' is equivalent to the standard '$\sim$' of strong Kleene logic. When viewed in this way, Visser's reconstruction of the argument seems correct. Similar reasons weigh against drawing the second conclusion.[30]

We are left with the third of my three primary options. I do not see any other viable options, but there is a reply I can offer on Field's behalf.

The reply questions the sense of adequacy involved. Everyone agrees that the strong Kleene material conditional is inadequate as a conditional, but there is less agreement, or even discussion, about what makes a conditional adequate in strong Kleene logic. Field must resort to using a revision-theoretic construction to introduce his conditional because

[30]What happens if step conditionals are added to the strong Kleene scheme instead of Field's conditional? If the result does not respond to Gupta's objections, then we have some evidence that the problems indicated above are due to the strong Kleene scheme, rather than just the particular conditional added. I thank Anil Gupta for raising this question.

the truth table for the successor stages of the construction is non-monotonic. If the truth table conditional were added to strong Kleene logic, then the Kripke construction would not reach a fixed-point. The fixed-point interpretation of the truth predicate is central to Field's view, so this route is unavailable to Field.

Field was not, as far as I can tell, concerned with responding to Gupta's objections to the fixed-point view of truth. His assessment of the adequacy for a conditional is different than that of Visser and Gupta. My assessment of Visser's idea implemented with Field's conditional will not be seen as a problem for a proponent of Field's conditional. The situation is different for someone that was moved by Gupta's objections to the strong Kleene fixed-point view of truth. My assessment shows that Field's conditional does not solve all of the major problems with the fixed-point view of truth. Insofar as one was unhappy with the minimal fixed-point view of truth, Field's conditional should not be seen as a major improvement.

6.3.2 Artifacts and deflationism

Field's conditional, and his theory of truth, are defined with respect to a single revision sequence. This revision sequence uses a constant limit rule: all unstable elements are assigned **n**. The theory of truth is the set of sentences assigned the value **t** at the least fixed-point over a reflection ordinal. These three things, the constant limit rule, the use of minimal fixed-points, and the use of reflection ordinals, create *artifacts* in the theory, erroneous verdicts based solely on *ad hoc* features of the formal construction. I will argue that these artifacts are a problem for Field's deflationary stance on truth.[31]

Field is a *deflationist* about truth. He thinks that all there is to the content of truth is captured by the Intersubstitutivity Principle. I submit that a consequence of this deflationism is the following. If we are given a simple entailment between two sentences involving truth and nothing else problematic, we should be able to explain its validity or invalidity by appeal to just the Intersubstitutivity Principle and the logic.

In addition to being a deflationist about truth, Field is an *instrumentalist* about model

[31]My use of the term "artifact" for the phenomenon to be discussed is primarily based on the usage in Yaqūb (1993, Ch. 3). The first example of this usage that I have been able to find is Belnap (1982b, 107).

theory. Field denies that the models he uses in defining his theory of truth have any philosophical significance or play any explanatory role. They are useful merely in proving consistency results and giving a feel for the valid arguments of the logic.

One of the roles a theory of truth has is that of telling us what arguments involving the truth predicate are valid. It is, I think, reasonable to criticize or reject a theory on the basis of its providing incorrect verdicts on many arguments. One of the sources of criticism stems from the presence of artifacts in the theory. Technical features of models, such as the minimal fixed-point, result in artifacts in the theory that provide incorrect results, such as the truth-teller entailing the liar. Belnap says, "[The Grand Loop] is an artifact of the construction, due entirely to the fact that the same Bootstrapper is used for each and every limit stage."[32] Field's construction enters into a Grand Loop, a cycle of hypotheses, for the very reason Belnap points out. Belnap continues, saying, "I think the Grand Loop is an artifact created by an *ad hoc* decision to adopt always a Constant Bootstrapping Policy where no such constancy is called for."[33]

Belnap criticized the constant limit rules of Herzberger and Gupta by pointing out strange results that one gets from constant limit rules. One such is the stability of the material equivalence between two distinct liar sentences. Yaqūb (1993) criticizes the revision theories proposed in Gupta (1982) and Belnap (1982b) on a similar basis. Yaqub argues that the revision theories contain too many artifacts that result in incorrect verdicts about the logical relations between sentences involving the truth predicate.

Given a rich enough ground language, one would expect that there will be some artifacts of the model that result in verdicts about entailments that are odd or counter-intuitive. Whether these can support a criticism of a theory depends on how widespread they are and how implausible they are. Field's theory produces too many artifacts and renders implausible assessments of them, and so it is subject to criticism on that basis.

In §6.3.1, I defined a sequence of iterated Curry sentences (Definition 25). We can, of course, define other sequences of iterated Curry sentences by changing false sentence in the inner-most consequent. As before, each of these sentences has a simple revision pattern. We

[32] Belnap (1982b, 107)
[33] Belnap (1982b, 107)

can define an analogous sequence, the iterated determinate liar.

Definition 26 (Iterated determinate liar)**.** *For each $n \geq 1$, d_n names the following sentence.*

$$\sim D^n T d_n$$

The determinate liars also have a simple pattern of revision. For each n, the pattern of revision for Td_n is the same as for Tc_n, except that where the latter has **f**, the former has **n**.

The iterated Curry sentences are built up out of a name, the truth predicate, arrows, and some false sentence. The false sentence can either be a falsity constant or a contradiction made of sentences from the base language. The iterated Curry sentences will receive the ultimate value **n** in all models of Field's theory. The iterated determinate liars are constructed from a name, the truth predicate, negation, conjunction, and the arrow, so they involve even less non-logical material. They will also receive the ultimate value **n** in all models of Field's theory. So, for all k, we will have the following entailments, for arbitrary A.

- $Tc_k \models A$
- $Td_k \models A$

This is somewhat expected, given that the consequence relation is defined as preservation of ultimate value **t**. According to Field's theory, a liar sentence entails everything. Let a name $\sim Ta$. Then the theory entails the following, for arbitrary A.

$$Ta \models A$$

Liar sentences are simply inconsistent according to Field's theory.

I want to emphasize that the iterated Curry and determinate liars are primarily constructed from logical constants, things that do not change their interpretation between models of the theory. These sentences are ones that are evaluated in the same way in all models of the theory, so their strange behavior can be pinned on the interpretation of the truth predicate and the conditional.

The strange behavior of the iterated Curry and determinate liars can be brought out by considering *conditionals* in which they feature as antecedents, rather than arguments in which they feature as premises. A simple calculation shows that $Tc_2 \rightarrow Tc_4$ is valid. In fact, for any even $k > 0$, $Tc_2 \rightarrow Tc_k$ is valid.

Proposition 4. *For $n, m > 0$, if $n \equiv 0(mod\ m)$, then $Tc_m \rightarrow Tc_n$ is valid.*

A similar point holds for the iterated determinate liars.

Proposition 5. *For $n, m > 0$, if $n \equiv 0(mod\ m)$, then $Td_m \rightarrow Td_n$ is valid.*

Let us say that conditionals are the *arrow correlates* of the entailment statements that are formed by replacing the main arrow of the former with a turnstile, '$\models$'. Field's conditional obeys modus ponens, so if we have a valid conditional, then corresponding entailment statement will be true. There are then many valid conditionals that have distinct iterated Curry sentences in their antecedents and consequents. Note, however, that not all entailments are reflected by arrow correlates, as illustrated the following.

(1) $Tc_4 \models Tc_2$

(2) $\not\models Tc_4 \rightarrow Tc_2$

(3) $\not\models \sim(Tc_4 \rightarrow Tc_2)$

While (1) is a true entailment statement, neither its arrow correlate nor the negation of its arrow correlate is valid. This is illustrated by the following table.

	...	k	k+1	k+2	k+3	k+4	...	ω
Tc_2		**t**	**f**	**t**	**f**	**t**		**n**
Tc_4		**t**	**t**	**t**	**f**	**t**		**n**
$Tc_4 \rightarrow Tc_2$		**t**	**t**	**f**	**t**	**t**		**n**
$\sim(Tc_4 \rightarrow Tc_2)$		**f**	**t**	**t**	**f**	**t**		**n**

Both $Tc_4 \rightarrow Tc_2$ and $\sim(Tc_4 \rightarrow Tc_2)$ will have ultimate value **n**.

In addition to many valid conditionals with distinct iterated Curry and determinate liars, there are also many valid conditionals containing iterated Curry sentences in their antecedents and iterated determinate liars in their consequents.

Proposition 6. *For $n, m > 0$, if $n \equiv 0(mod\ m)$, then $Tc_m \rightarrow Td_n$ is valid.*

The result of switching the position and subscript of the iterated Curry and the iterated determinate liar in the proposition, $Td_m \rightarrow Tc_n$, will not be valid. The reason is that the iterated determinate liars cycle between values **t** and **n**, whereas the iterated Curry sentences

cycle between **t** and **f**. There will be stages at which the antecedent will have value **n** while the consequent has value **f**, so at the next stage the conditional will have value **f**.

The validity of these conditionals is, I think, fairly described as an artifact of the revision process defining Field's conditional. Let us say call the valid conditionals highlighted in propositions (4), (5), and (6) *artifactual conditionals*. There does not seem to be any way to prove one of the artifactual conditionals on the basis of the propositional logic with the conditional and the T-sentences. This can even be supplemented with the syntactic theory. There still appears to be no way to prove the artifactual sentences.

Field could respond to this by saying that he takes the artifactual sentences as additional axioms. This is a poor response, because taking on additional axioms for truth gives up on deflationism. It would be to accept that there is more to truth than the Intersubstitutivity Principle, namely these additional axioms. Accepting deflationism while maintaining that there is no way to derive the artifactual conditionals on the basis of logic and the Intersubstitutivity Principle is not coherent.

There does not seem to be any way for Field to justify the validity of the artifactual conditionals without appeal to the revision sequences for his conditional. The revision sequences allow us to see why the artifactual conditionals come out valid, but they do not provide any reason for thinking that those are good evaluations.

One of the key features of Field's models is that they have many truth values. Suppose that the period of the revision process between reflection ordinals is Σ. We can view each Σ-long sequence from $\{\mathbf{t}, \mathbf{n}, \mathbf{f}\}$ that a sentence can take as a possible value. These values are only partially ordered, not linearly ordered. Despite the multitude of values, the definition of the conditional imposes such simple patterns on paradoxical sentences, that the value assigned to a paradoxical sentence, such as an iterated Curry sentence, stands in the ordering relation to many other paradoxical sentences. When the ordering relation holds between the values assigned to sentences A and B, then $A \to B$ is true. Even though there are many values available in the model, Field's construction hardly exploits this fact. The result is that Field has a model theory for his theory of truth that is cumbersome and yields many counterintuitive and incorrect verdicts.

There are many artifactual sentences. These are a problem for Field's deflationism

because there does not seem to be any way to argue for them using Field's logic together with the Intersubstitutivity Principle. I suggested that Field could accept additional axioms for truth, but this response would be, I claimed, giving up on deflationism. The additional axioms would provide content to the notion of truth beyond the Intersubstitutivity Principle. According to Field's deflationism, the Intersubstitutivity Principle exhausts the content of truth. No other principles are needed to codify the logic of truth. Yet, the basic logic together with the Intersubstitutivity Principle do not appear to allow us to argue for the artifactual sentences. Therefore, Field's narrow theory of truth is inconsistent with his philosophical views about truth, his wide theory.

Field does not have an easy way out of this criticism. A simple adjustment to the truth table used in the revision rule for his conditional will not eliminate all, or even most, of the artifactual sentences. Adjusting the limit policy to be a non-constant policy will eliminate many more of the artifactual sentences, but this will come at a steep cost. Field will lose the good behavior of the determinateness operator, particularly iterations of it. I do not think that Field would want to alter the behavior of the determinateness operator, because he views it as one of the key features of his theory.

In his response to Yablo, Field proposes some other ways of modifying the revision rule for his conditional.[34] These modifications quantify over fixed-points at each stage. Note that the artifactual sentences I highlight do not depend on revising according to the least fixed-point at each stage. Changing which fixed-point is used, or quantifying over many of them, will not eliminate my examples.

There are likely other ways of modifying the revision rule that will eliminate some of the artifactual sentences. I expect that other modifications will increase the complexity of the revision process, which will in turn make the logic and the semantics more unwieldy. Increased complexity would, I think, be a large detriment to Field's theory.

There is one more response to consider on Field's behalf. In §6.2.1, I quoted Field as saying that real validity is some effectively generable subset of the validity relation he defined. This response, the axiomatization response, says that the artifacts to which I point are not really valid, so my criticism has no force.

[34] See Field (2008, 17.5) for the details.

This response, while initially appealing, will not work. To begin, note that Field needs the effectively generable subset to be nice to do the job he wants, since the empty set is an effectively generable subset that will not work as an axiomatization of a logic.[35] There is then a question of whether such a subset exists.

The potential problem for Field's axiomatization response is that it risks losing the nice behavior of the determinateness operator. The reason is that the models ensure that for any defective sentence, there is some α for which the sentence is not α-determinately true. This is touted, rightly I think, as a positive feature of the theory.[36] It is hard to see how an axiomatization of the logic together with the Intersubstitutivity Principle will ensure that defective sentences will be categorized as such.[37] For some sentences, such as liars and determinate liars, one could add in axioms ensuring that they are indeterminate, in some sense. For any of the array of iterated Curry sentences, let alone paradoxical sentences that use quantifiers, it is difficult to imagine how one could capture their defectiveness axiomatically.

The axiomatization response will not work. Although it would provide a way to fend off my criticism based on the presence of artifacts, assuming an adequate axiomatization could be given, it threatens to undermine a key feature of Field's view. Field could adopt the axiomatization response, but it would require reevaluating the new overall theory, which requires waiting upon the axiomatization.

6.3.3 Truth-preservation

There is one other negative feature of Field's theory of truth that I will highlight. In Field (2006b) and Field (2008), Field stresses the importance of having a theory of truth accept that the rules under which it is closed preserve truth.[38] In the terminology of §6.3.2, the

[35] I thank Anil Gupta for pointing this out.

[36] See Field (2008, 276), for example.

[37] We will set aside the empirical case and restrict attention, as Field does in Field (2008), to arithmetic or set theory.

[38] The earlier of the two sources places greater importance on this, although the later source does use failure of the truth-preservation as an objection against many theories. See Field (2008, Ch. 26) for the use of failure of truth-preservation as a primary criticism of a theory.

truth-preservation claim for the rule

$$A_1, \ldots, A_n \models B$$

is the validity of its arrow correlate,

$$\models T(\ulcorner A_1 \urcorner) \,\&\, \ldots \,\&\, T(\ulcorner A_n \urcorner) \rightarrow T(\ulcorner B \urcorner) \,.$$

Field argues that his theory of truth entails the truth-preservation claims for the truth rules and other rules under which his theory is closed. He also argues that the theory does not entail any counterexamples or disjunctions of counterexamples to any particular instance of a rule.

Field does accept the *ex falso* rule.

$$A, {\sim}A \models B$$

The arrow correlate of this is not valid. In fact, there are instances where the arrow correlate is contravalid. Let Ta be a liar sentence and substitute that for A and substitute a ground language falsehood or contradiction for B. We have the following.

$$\models {\sim}(Ta \,\&\, {\sim}Ta \rightarrow 0 = 1)$$

Since truth obeys the Intersubstitutivity Principle for Field, this is equivalent to the negation of the truth-preservation claim for *ex falso*. [39] The problem is that Field's theory entails the rejection of the *ex falso* rule that Field wants for his logic. By Field's own lights, this is a large flaw in a theory of truth.

Field will be unable to recover truth-preservation for *ex falso*, for a simple reason. Let us fix the sentence substituted for B as a ground language contradiction, $\bot$. The arrow correlate then becomes

$$Ta \,\&\, {\sim}Ta \rightarrow \bot.$$

[39] Rather than the arrow correlate of *ex falso*, the definition of Field's conditional ensures that $A \,\&\, {\sim}A \rightarrow B \vee {\sim}B$ is valid. The rule to which this is the arrow correlate, however, is a weakened *ex falso* rule under which Field's theory is closed.

Contraposing this yields

$$\top \rightarrow \sim(Ta \,\&\, \sim Ta),$$

which is in turn equivalent to

$$\top \rightarrow Ta \vee \sim Ta.$$

Thus, the truth-preservation claim for *ex falso* results in excluded middle for all sentences. This would trivialize Field's theory as its response to the paradoxes is to reject excluded middle except when it is posited as an axiom for certain safe vocabulary. The above reasoning uses full contraposition, one of the de Morgan's laws, and double negation elimination. None of these principles are negotiable for Field.

Field's theory of truth does not say that its rules are truth-preserving. In fact, it says that some of its rules do not preserve truth, *pace* Field.[40] By Field's own lights, this is a failing of his theory. It closes the philosophical gap Field sees between his theory and the so-called weakly classical theories such as revision theories. Additionally, it throws into relief the differences between the consequence relation of Field's logic and the logical behavior of the conditional.

There are two major points at which Field's view fails to meet my criteria. The first deals arguments and the second with neutrality. I will cover these points briefly.

As I have argued, Field's view gives incorrect verdicts on many arguments involving truth. In addition it is not clear what the logic is, which hinders the evaluation of arguments.

The main way in which Field's view runs afoul of neutrality is that it cannot be combined with certain logical resources in the base language. This is due to features of the underlying fixed-point view. Although Field's construction permits the addition of a new conditional that, in a sense, lacks the monotonicity property, the conditional is not present in the base language of the theory. Additionally, it is unclear how to combine Field's conditional with non-classical base languages.[41] One option is to use one of the non-designated values in the way that Field uses **n**. Another option is to add a new semantic value, corresponding to **n**, below all the other values in the ordering of semantic values. Each of these seems to be in the spirit of Field's construction with little to choose between them.

[40] Field (2006b, 590-591)

[41] I owe this point to Anil Gupta.

I have presented two criticisms of Field's logic on the basis of its adequacy when applied to truth. I also showed how it does not respond to three of the four objections offered by Gupta to fixed-point theories of truth. I gave a brief summary of how Field's view violates two of the criteria argued for in chapters 1 and 2. I think these points together provide strong evidence that Field's theory should be rejected on philosophical grounds. I will now turn to the more technical portion of my discussion of Field's logic.

6.4 FIELD'S PROPOSITIONAL LOGIC

Field's canonical models build in complexity that is extraneous to the study of the propositional logic. Using a revision process, he obtains a complete lattice of values, the sequences of truth values that show up in the loop of the revision process. The values available depend on the resources in the language that can be used to define pathological sentences. There is no reason for the study of the logic to be restricted to the sequences that can be defined in a given language.[42] In this section, I will develop a framework that gets around this shortcoming.

An idea similar to my framework is present in Field (2008). Field talks about the sequences of values in the loops of the revision process forming a lattice. Field's discussion is restricted to the lattices that one obtains via the revision process over models of set theory or arithmetic. In a similar vein, Priest (2010) makes a connection between Field's lattices and the lattices used in algebraic semantics for weak relevance logics. Priest's discussion undercuts the role of the revision process in the semantics, since the lattice of values is all that is needed for the connection that Priest draws between Field's approach and relevance logic. Additionally, the result crucial to Priest's discussion depends on the presence of a fusion connective in the language.[43] Field's logic lacks a fusion connective and there are

[42] If one wants to incorporate more self-reference without concern for definability in arithmetic or set theory, one could use Visser's stipulations, as found in Visser (2004). There will, however, be questions about what patterns of truth values can be produced with stipulations. I owe this suggestion to Anil Gupta.

[43] Fusion is a connective found most often in discussions of relevance logic. It is an "intensional conjunction" and it residuates relevance conditionals. See Anderson and Belnap (1975), among other things, for discussion.

some difficulties with adding such a connective to the logic with Field's conditional.[44] My framework, unlike Priest's, builds in aspects of the revision process in generating semantic values.

The point of using this general framework is to see what the propositional logic of Field's conditional is. We have some idea of the laws and inferences validated by his canonical models, but it is not clear which of those are accidentally validated, so to speak, due to restrictions arising from certain *ad hoc* features of the construction. Because some of the laws and inferences may be invalidated by our more general framework, we will have to explore several options for the semantics to see which captures the most attractive logic or the one most faithful to Field's views. Based on Field (2008), some of the inferences, such as *modus ponens*, should be valid and some should definitely be invalid, such as the rule form of contraction. Some are less certain, and those will depend on the framework.

I will present the definitions needed for the general study of Field's conditional. I will use the notation $x \in [\alpha, \beta)$ to abbreviate $\alpha \leq x < \beta$.

Definition 27 (Values)**.** $Sequences(\alpha) = \{f : f \in \alpha \mapsto \{\boldsymbol{t}, \boldsymbol{n}, \boldsymbol{f}\}\}$

Let $Values(\gamma)$ *be the set of all elements* $x \in Sequences(\gamma)$ *satisfying, for all limit ordinals* $\lambda \leq \gamma$*, the following two conditions.*

1. $\forall \alpha < \lambda \exists \beta \in [\alpha, \lambda) x_\beta \neq \boldsymbol{t} \Rightarrow x_\lambda \neq \boldsymbol{t}$
2. $\forall \alpha < \lambda \exists \beta \in [\alpha, \lambda) x_\beta \neq \boldsymbol{f} \Rightarrow x_\lambda \neq \boldsymbol{f}$

The set $Values(\alpha)$ is the set of semantic values of a given ordinal length. I will exclusively use limit ordinals. Elements from this set will be used to interpret sentence letters.

Some comments about the set $Values(\alpha)$ are in order. The conditions on the sequences in $Values(\alpha)$ are weaker than the conditions of coherence on revision sequences. Revision theory requires that if a sentence has stabilized, say, at **t** going up to a limit stage, then its value at the limit stage will be **t**. According to the conditions given, the value of such a sentence at the limit may be **n**. The reason is that some sentences, such as $Tc \vee \sim Tc$, where c is a Curry sentence, may be stably **t** going up to a limit but **n** at the limit. Rather than use one set of values for atoms and another set for complex sentences, I have opted to use

[44] The point about fusion and its connection to Priest's claim was brought to my attention by Greg Restall.

the weaker condition for values for all sentences.

I will frequently talk about the length λ of the set of values, which means that I am working with $Values(\lambda)$. When it is clear what the ordinal length is, I will talk about $Values$ and the other notions with the ordinal suppressed.

Next, the operations corresponding to the logical connectives are defined pointwise on $Values(\lambda)$ as follows.

Definition 28 (Operations)**.** *Suppose* $x, y \in Values(\lambda)$.

- $x \rightarrow y = \langle \boldsymbol{n} \rangle^\frown \langle x_\alpha \rightarrow y_\alpha :$ *for all* $\alpha < \lambda \rangle$, *where '*$\rightarrow$*' is the truth table operator for Field's arrow on successor stages and the limit rule on limit stages.*
- $x \vee y = \langle x_\alpha \vee y_\alpha :$ *for all* $\alpha < \lambda \rangle$, *where '*$\vee$*' is the strong Kleene disjunction.*
- $x \wedge y = \langle x_\alpha \wedge y_\alpha :$ *for all* $\alpha < \lambda \rangle$, *where '*$\wedge$*' is the strong Kleene conjunction.*
- $\sim x = \langle \sim x_\alpha :$ *for all* $\alpha < \lambda \rangle$, *where '*$\sim$*' is the strong Kleene negation.*

Before proceeding, it is worth observing a couple of facts about the definitions given so far. Suppose that $x, y \in Values(\lambda)$. Going up to some successor limit β, $x \wedge y$ may have the value **f** at all successor stages, yet have the value **n** at β. This can happen when x alternates in the pattern **tf** over successor stages while y alternates in the pattern **ft**. On Field's approach, logically complex sentences whose main connectives are conjunction or disjunction need not have their stable values at limits. This is in contrast with the standard revision theory, for which all sentences retain their stable values at limits. This distinction is important for Field's view. The following definition captures the two parts of this idea.

Definition 29 (TrueOr, FalseAnd)**.** *For* $v \in V(\lambda)$, *let* $TrueOr(v, \lambda)$ *be defined as*

$$\forall A, B (\textit{if } \exists \alpha < \lambda \forall \beta \in [\alpha, \lambda)\ v(A \vee B)_\beta = \boldsymbol{t}, \textit{then}$$

$$\exists \alpha < \lambda \forall \beta \in [\alpha, \lambda)\ v(A)_\beta = \boldsymbol{t} \textit{ or } \exists \alpha < \lambda \forall \beta \in [\alpha, \lambda)\ v(B)_\beta = \boldsymbol{t}).$$

For $v \in V(\lambda)$, *let* $FalseAnd(v, \lambda)$ *be defined as*

$$\forall A, B (\textit{if } \exists \alpha < \lambda \forall \beta \in [\alpha, \lambda)\ v(A \,\&\, B)_\beta = \boldsymbol{f}, \textit{then}$$

$$\exists \alpha < \lambda \forall \beta \in [\alpha, \lambda)\ v(A)_\beta = \boldsymbol{f} \textit{ or } \exists \alpha < \lambda \forall \beta \in [\alpha, \lambda)\ v(B)_\beta = \boldsymbol{f}).$$

These predicates say that if a disjunction[conjunction] stabilizes at **t**[**f**] then one of the disjuncts[conjuncts] has also stabilized at that value. These are properties that are needed for conjunction and disjunction to behave in an appropriately strong Kleene-like fashion.

Definition 30 (K(λ)-valuation)**.** *A* K(λ)-valuation *is any valuation v for λ-sequences that satisfies $FalseAnd(v,\lambda)$ & $TrueOr(v,\lambda)$.*

$\mathcal{K}(\lambda) = \{v \in V(\lambda) : v$ *is a K(λ)-valuation*$\}$.

We will generally suppress the ordinal when discussing $K(\lambda)$-valuations, since a particular ordinal-length will be supplied by the context.

Next, we must define the ordering on the semantic values. The ordering $\preceq$ on the set $Values(\lambda)$ is defined as eventually greater than or equal.

Definition 31 (Ordering)**.**

$$\forall x, y \in Values(\lambda)(x \preceq y \iff \exists \alpha < \lambda \, \forall \beta \in [\alpha, \lambda) \; x_\beta \leq y_\beta)$$

The ordering $\preceq$ is a pre-order. There are many order-equivalent elements. The $\preceq$-equivalence, $\approx$, is an equivalence relation, so it can be used to construct quotient spaces. The spaces in which we are ultimately interested are the quotients of $Values(\lambda)$, for different ordinals λ, and the orderings on the quotient generated by $\preceq$. I will use '$\preceq$' for the quotient ordering as well. For each quotient of $Values(\lambda)$, the ordering on the quotient is defined as follows.

$$[x] \preceq [y] \text{ iff } x \preceq y$$

The generated ordering is a partial order. The operations will be defined on the elements of the quotient space in the standard way.

- $[x] \circ [y] = [x \circ y]$, for $\circ \in \{\rightarrow, \wedge, \vee\}$
- $\sim[x] = [\sim x]$

My focus will be on the quotient spaces and their orderings. Later, instead of using the standard notation, $[x]$, for elements of quotient spaces, I will use x. I will use "$Values(\lambda)$" to denote the quotient space obtained from $Values(\lambda)$. Later, questions regarding K-valuations will be important. We will say that an interpretation for the quotient space is a K-valuation

if for each pair of sentences A and B, every triple of sequences from $v(A \vee B), v(A)$, and $v(B)$ satisfies the $TrueOr$ and $FalseAnd$ conditions.

I will use $\overline{\mathbf{t}}, \overline{\mathbf{n}}$, and $\overline{\mathbf{f}}$ as constants for the λ-long sequences that eventually stabilize at the strong Kleene values $\mathbf{t}, \mathbf{n}$, and $\mathbf{f}$, respectively.

The pointwise meet and join are meet and join for the ordering on the quotient spaces $Values(\lambda)$. Suppose that $[x] \preceq [y]$ and $[x] \preceq [z]$. Then $\exists \alpha < \lambda \, \forall \beta \in [\alpha, \lambda) \; x_\beta \leq y_\beta$ and $\exists \alpha' < \lambda \, \forall \beta' \in [\alpha', \lambda) \; x_{\beta'} \leq z_{\beta'}$. Let γ be the larger of α and α'. Then $\exists \gamma < \lambda \, \forall \beta \in [\gamma, \lambda) \; x_\beta \leq y \wedge z_\beta$, so $x \preceq y \wedge z$, so $[x] \preceq [y \wedge z]$. Next, we show that $[x \wedge y] \preceq x$. $[x \wedge y] \preceq x$ is equivalent to $\exists \alpha < \lambda \, \forall \beta \in [\alpha, \lambda) \; x \wedge y_\beta \leq x_\beta$, which is clearly true by the definition of meet at each stage. The argument for $[x \wedge y] \preceq [y]$ is similar.

One thing to notice is that the quotient space $Values(\alpha)$ is a lattice, but it is not a complete lattice. Suppose that the set of values $X = \{x^i : i \in \omega\}$ is such that for each x^i, limit $\lambda < \alpha$, and natural numbers $k, m > 0$, $x^i_{\lambda+k} = \mathbf{n}$ for $k \leq i$ and $x^i_{\lambda+m} = \mathbf{t}$ for $m > i$. Suppose also that each x^i obeys the revision condition, so they are $\mathbf{t}$ at limits. They can take any value at stage 0. At each successor stage, the meet of X is $\mathbf{n}$, At limit stages, however, the meet is $\mathbf{t}$. Sequences with with that pattern of values are not in $Values(\alpha)$, by clause 1 of its definition.

I will now show that this framework captures some of the essential features of the semantics of the conditional. The following two theorems hold for sequences of any limit ordinal length.

Theorem 30. $x \rightarrow y = \overline{\mathbf{t}} \iff x \preceq y$

Proof. Both directions are immediate from definitions. □

We also get the following.

Theorem 31. *(1) If $x \preceq y$, then $z \rightarrow x \preceq z \rightarrow y$.*

(2) If $x \preceq y$, then $y \rightarrow z \preceq x \rightarrow z$.

(3) $\overline{\mathbf{t}} \rightarrow \overline{\mathbf{f}} = \overline{\mathbf{f}}$.

(4) If $x = \overline{\mathbf{t}}$, then $y \rightarrow x = \overline{\mathbf{t}}$.

Proof. For (1), suppose that $x \preceq y$. Assume that $z \to x \not\preceq z \to y$. Then $\forall\alpha < \lambda\ \exists\beta \in [\alpha, \lambda)\ (z \to x)_\beta \not\leq (z \to y)_\beta$. Let α be greater than the point at which $x \to y$ stabilizes to $\mathbf{t}$. Fix β as the least ordinal greater than α, such that $(z \to x)_\beta \not\leq (z \to y)_\beta$. There are now two cases, depending on whether β is a successor or limit, of which we will present only the successor case.

Case: $\beta = \gamma + 1$. $z \to x_\beta = \begin{cases} \mathbf{t} & z_\gamma \leq x_\gamma \\ \mathbf{f} & \text{otherwise} \end{cases}$

Either $z_\gamma \leq x_\gamma$ or $z_\gamma > x_\gamma$.

Subcase: Suppose $z_\gamma \leq x_\gamma$. Then, since $\alpha \leq \gamma$, $x_\gamma \leq y_\gamma$, so $z_\gamma \leq y_\gamma$, so $(z \to y)_\beta = \mathbf{t}$. Therefore $(z \to x)_\beta \leq (z \to y)_\beta$ This contradicts the assumption.

Subcase: Suppose $z_\gamma > x_\gamma$. Then $(z \to x)_\beta = \mathbf{f}$, so $(z \to x)_\beta \leq (z \to y)_\beta$, which contradicts the assumption.

In both subcases we reached a contradiction.

Both cases result in contradiction, so we conclude that if $x \preceq y$, then $z \to x \preceq z \to y$.

The argument for (2) is similar.

For (3), note that for $\forall\beta \in [0, \lambda)\overline{\mathbf{t}}_\beta > \overline{\mathbf{f}}_\beta$, so $\forall\beta \in [1, \lambda)(\overline{\mathbf{t}} \to \overline{\mathbf{f}})_\beta = \mathbf{f}$.

For (4), note that if the consequent of an arrow is $\mathbf{t}$, then the arrow is $\mathbf{t}$ as well. It follows that if the consequent is stably $\mathbf{t}$, then the arrow is as well. □

The preceding two theorems establish that the semantic framework captures the minimal relations needed to study Field's conditional. There are issues that arise when using this framework to study the conditional in the context of the other propositional connectives. The rest of this section will be devoted to sorting out those issues. Before getting to those issues, I will need a few more definitions.

We will use a fixed language L with a countably infinite set of proposition letters and complex sentences formed from the atoms, the strong Kleene connectives and the conditional.

Definition 32 (Interpretations)**.** *Let $I(\alpha)$ be the set of functions from proposition letters of L to $Values(\alpha)$.*

Let $V(\alpha)$ be the set of valuations induced by $I(\alpha)$.

We need to specify the how logically complex sentences are evaluated. Suppose $v(A) = x$ and $v(B) = y$.

- $v(\sim A) = \sim x$
- $v(A \,\&\, B) = x \wedge y$
- $v(A \vee B) = x \vee y$
- $v(A \rightarrow B) = x \rightarrow y$

Definition 33 (Ultimate value)**.** *The* ultimate value of a sentence A*, where A is an atom or has the form $B \rightarrow C$, with respect to a valuation v is $\boldsymbol{t}[\boldsymbol{f}]$ if the sequence v assigns A is eventually stably $\boldsymbol{t}[\boldsymbol{f}]$. Otherwise, the ultimate value is $\boldsymbol{n}$.*

If A has the form $\sim B$, $B \,\&\, C$, or $B \vee C$, then the ultimate value is the recursively defined on the ultimate values of B alone or B and C, respectively, using the strong Kleene truth tables for $\sim, \wedge,$ and $\vee$.

If A has the ultimate value of $\boldsymbol{t}[/\boldsymbol{n}/\boldsymbol{f}]$ on a valuation v, we will write it as $||A||_v = \boldsymbol{t}[/\boldsymbol{n}/\boldsymbol{f}]$.

Definition 34 ($\models$)**.** *A set of sentences Γ* λ-entails *a sentence A or the argument from Γ to A is* λ-valid*, in symbols $\Gamma \models_\lambda A$, when all valuations in $V(\lambda)$ that assign ultimate value $\boldsymbol{t}$ to all the sentences in Γ also assign ultimate value $\boldsymbol{t}$ to A.*

In what follows, I will suppress the ordinal parameter of entailment.

The values $v(A)$ assigned to a sentence A by v are equivalence classes of sequences. For each pair of sequences y and z in a equivalence class $[x]$, there is some point α after which $y_\alpha = z_\alpha$. In talking about the valuations that invalidate certain laws and inferences, it will be useful to specify certain sequences of the underlying set of values rather than the values of the quotient space. This is okay since each such sequence belongs to exactly one equivalence class. Therefore, I will use the notation $v(A)_\beta$ to refer to x_β, where x is a particular sequence in $v(A)$.

With the definitions in place, I can now investigate how the length of the values affects the logic. I will look at sequences of three lengths: ω, countable, uncountable. An important question regarding each of these lengths is whether the stable value of a sentence must coincide with its ultimate value.

6.4.1 Short sequences

We will begin with the shortest sequences, those of length ω. They may be adequate for capturing the logic of the pure arrow fragment. Based on the theorems of the previous section, prefixing and suffixing are both valid. Additionally, $A \to A$ is valid.

The ω-long sequences also invalidate the laws and entailments that Field wants to invalidate, including the following.

1. $\not\models A \to .B \to A$
2. $\not\models A \to B \to .B \to C \to .A \to C$
3. $A \to .A \to B \not\models A \to B$
4. $A \to .B \to C \not\models B \to .A \to C$

The first is invalidated by a valuation that assigns stably **n** to A and stably **t** to B. The second is invalidated by the valuation that assigns stably **t** to A and assigns to both B and C the sequence made of **n** followed by **tftftftf**.... The third is invalidated by a valuation that assigns stably **f** to B and **ntftftftf**... to A. The fourth is invalidated by assigning stably **f** to C, and **nff** followed by **tff**... to A and **ntf** followed by **fttttf**... to B.

The use of ω sequences to interpret the atomic sentences is inadequate for the full language of Field's logic, that is, the strong Kleene connectives and the conditional, because then *modus ponens* will be invalid.

	0	1	2	3	4	5	...	ω
A	**t**	**t**	**t**	**f**	**t**	**f**		**t**
B	**t**	**f**	**t**	**f**	**t**	**f**		**n**
C	**f**	**t**	**f**	**t**	**f**	**t**		**n**
$B \vee C$	**t**	**t**	**t**	**t**	**t**	**t**		**n**
$A \to B \vee C$	**n**	**t**	**t**	**t**	**t**	**t**		**t**

In this table, the ω column corresponds to the ultimate value. Both A and $A \to B \vee C$ have ultimate value **t**, but $B \vee C$ does not. $B \vee C$ stabilizes at **t**, but its stable value does not correspond to its ultimate value.

The ω sequences also invalidate an argument that Field wants to come out valid.

$$\sim(A \to B) \not\models A \vee \sim B$$

The interpretation on which A oscillates in the pattern **tn** and B oscillates in the pattern **nf** makes $\sim(A \to B)$ eventually stably **t**, and so its ultimate value is **t**. The ultimate value of $A \vee \sim B$ on this interpretation is **n** as the ultimate value of both A and $\sim B$ is **n**. It is not too hard to see that the corresponding arrow form,

$$\sim(A \to B) \to A \vee \sim B,$$

is valid.

From these examples, we can see that the semantic evaluation of a disjunction in the scope of an arrow is, for the purposes of determining validity, importantly different from the evaluation of a free-standing disjunction.[45] If a disjunction is a premiss, one of the disjuncts must have ultimate value **t**. If a disjunction is, say, the consequent of an arrow with a stably **t** antecedent, then the eventual stability of the disjunction at **t** is all that is needed for the ultimate value of the conditional to be **t**. Eventual stability of a disjunction does not determine the ultimate value of the disjunction in general. In K-valuations, the two notions do coincide.

As the earlier examples show, we cannot use sequences that are ω-long as the semantic values if there are connectives in the language besides $\to$. In this case, restricting the valuations to those in $\mathcal{K}(\omega)$ will not help, as there is not sufficient variation in ω-long sequences in $\mathcal{K}(\omega)$. At all successor stages, $(A \to B) \vee \sim(A \to B)$ gets the value **t**. In particular, for any two atomic sentences p and q, $(p \to q) \vee \sim(p \to q)$ will be stably **t**. A K-valuation must make either $p \to q$ or $\sim(p \to q)$ eventually stably **t** as well as one of $q \to p$ and $\sim(q \to p)$.

If we require that for each conditional, either it or its negation must be stably true, then we lose many of the valuations needed to invalidate some of the undesired laws and inferences. For example, the counterexample to the contraction inference given above makes neither $A \to B$ nor $\sim(A \to B)$ stably **t**. The restrictions of using only valuations in $\mathcal{K}(\omega)$ are too great.

[45] The same point holds for false conjunctions.

The solution to the problem of not having enough "good" valuations is to use longer sequences in order to obtain more limit ordinals.

6.4.2 Longer sequences

Some of problems with ω-long K-valuations arose because there were not sufficient limit ordinals. Those problems will arise again if a successor limit ordinal is used as the length of sequences. There are several natural options for ordinals that are limits of limit ordinals and countable, such as ω^2, ω^ω, and ϵ_0.

Many of the features of the ω-long sequences are maintained with the move to longer sequences. Some of the flaws are removed. There is now sufficient variation in the valuations in λ to invalidate contraction as well as not requiring that conditionals involving atoms come out as stably **t** or stably **f**.

One feature of using $V(\lambda)$, where λ is countable and a limit of limit ordinals, is that the effects of the limit stages can be clearly seen. Some conjunctive strengthenings of inferences, such as

$$A \rightarrow B \,\&\, B \rightarrow C \rightarrow .A \rightarrow C,$$

are bound to stabilize at **t** over successor stages. If the conjuncts of the antecedent alternate so that at least one is **f** at each successor stage but neither is stably **f**, then at limits both conjuncts will be **n**, so the antecedent will be **n**.

As in the case of ω, $\mathcal{K}(\lambda)$ is a proper subset of $V(\lambda)$. Restricting to $\mathcal{K}(\lambda)$ gives us the validity of *modus ponens*. If we use $V(\lambda)$ instead of $\mathcal{K}(\lambda)$ as the set of valuations, then *modus ponens* will be invalid. Here is an example for $V(\omega^2)$, which is similar to the counterexample to *modus ponens* for $V(\omega)$.

$$A, A \rightarrow B \vee C \not\models B \vee C$$

To see this, let $v \in V$ be such that $||A||_v = \mathbf{t}$, the values $v(B)$ and $v(C)$ are each **n** at limits, and for all n and m,

- $v(B)_{\omega*2n+(m+1)} = \mathbf{t}$
- $v(C)_{\omega*2n+(m+1)} = \mathbf{f}$
- $v(C)_{\omega*2n+\omega+(m+1)} = \mathbf{t}$

- $v(B)_{\omega*2n+\omega+(m+1)} = \mathbf{f}$

The result is that on v, $B \vee C$ is stably $\mathbf{t}$ while neither B nor C is stably $\mathbf{t}$. We know that the $||A||_v = ||A \rightarrow B \vee C||_v = \mathbf{t}$, but $||B \vee C||_v = \mathbf{n}$.

If we had defined ultimate value to be eventually stably $\mathbf{t}[\mathbf{f}]$ or unstable for all connectives, then disjunction would behave differently. We could, alternately, obtain a different notion of entailment by using preservation of stably $\mathbf{t}$ rather than preservation of ultimate value $\mathbf{t}$.

A consequence of this example is that the $\models$ relation is not closed under the substitution of sentences for atoms. The entailment, $p, p \rightarrow q \models q$, holds, where p and q are atoms, but it fails when substituting a disjunction for q. In addition to the invalidity of *modus ponens*, the failure of substitution is another strike against using countable sequences for the definition of consequence.

A natural idea is to restrict the set of valuations to just those in $\mathcal{K}(\lambda)$ rather than $V(\lambda)$ in the definition of consequence, but there are two problems with this idea. The first is that it requires the extra step of verifying that a putative counterexample valuation is actually a K-valuation. The second is that the restriction requires ignoring some valuations. One of the motivations for adopting this framework was to eliminate *ad hoc* restrictions, so it seems that we are putting them back in. The solution is to find ordinals λ such that $V(\lambda) = \mathcal{K}(\lambda)$

6.4.3 Even longer sequences

One idea to find ordinals λ such that $V(\lambda) = \mathcal{K}(\lambda)$ would be to try to figure out what ordinals count as reflection ordinals for our framework. Reflection ordinals are limit ordinals where the sentences that are stable at that ordinal are all and only the sentences stable in the revision process carried on throughout all the ordinals, and similarly for the unstable sentences. The problem with this idea is that in our framework, given any ordinal, a sentence may be stable up to that ordinal but unstable afterwards. The conditions on sequences in $Values(\lambda)$ reflect partial coherence with a revision process but not the other restrictions built into a revision operator. Reflection ordinals play an important role in Field (2008). The canonical models of Field's theory of truth are the fixed-points over reflection ordinal

stages of the model construction.[46] The main purpose of the reflection ordinals for Field is to ensure that disjunction and conjunction behave in a strong Kleene-like fashion, which is to say that reflection ordinals guarantees that he is using K-valuations.[47] The main question is whether there is any length of sequences for which all valuations are K-valuations. The answer is yes.

Theorem 32. *All valuations in $V(\omega_1)$ are K-valuations.*

Proof. Suppose that there is an ω_1-valuation v such that $v(A \vee B)$ eventually stabilizes to $\mathbf{t}$ but neither $v(A)$ nor $v(B)$ does. This means that

$$\exists \alpha < \omega_1 \, \forall \beta \in [\alpha, \omega_1) \; v(A \vee B)_\beta = \mathbf{t}$$

but

$$\forall \alpha' < \omega_1 \exists \gamma, \delta \in [\alpha + 1, \omega_1) v(A)_\gamma \neq \mathbf{t} \,\&\, v(B)_\delta \neq \mathbf{t}.$$

Let S and T be ω-long sequences of limit ordinals less than ω_1 that satisfy the following properties.

- $\forall n S_n < T_n < S_{n+1}$
- $S_0 > \Sigma$, where Σ is the stabilization point of $v(A \vee B)$.
- $v(A)_{S_n} \neq \mathbf{t}$ and $v(B)_{T_n} \neq \mathbf{t}$

There are such sequences by the assumption that $v(A) \neq \mathbf{t}$ and $v(B) \neq \mathbf{t}$ together with the use of $Values(\omega_1)$. Both $sup(S)$ and $sup(T)$ are limit ordinals less than ω_1, because all ω-long sequences of ordinals from ω_1 are bounded in ω_1.

Next, we claim that $sup(S) = sup(T)$. Assume $sup(S) \neq sup(T)$. Then either $sup(S) < sup(T)$ or $sup(S) > sup(T)$. Assume $sup(S) < sup(T)$. Then there is some n such that $T_n > sup(S)$, which is impossible because $S_{n+1} > T_n$, by definition. The case in which $sup(S) > sup(T)$ is similar. In both cases, we have a contradiction, so $sup(S) = sup(T)$.

Since $sup(S)$ is a limit ordinal, and $\forall \beta \exists \gamma \in [\beta, sup(S)) v(A)_\gamma \neq \mathbf{t}$, $v(A)_{sup(S)} \neq \mathbf{t}$. For similar reasons, $v(B)_{sup(S)} \neq \mathbf{t}$. But, then, $v(A \vee B)_{sup(S)} \neq \mathbf{t}$, from the definition of

[46] Field uses the term "admissible ordinal" for reflection ordinals.
[47] See the fundamental theorem of Field (2008, 257-258).

a valuation. This is a contradiction, since $sup(S) > \Sigma$, which was the ordinal at which $v(A \vee B)$ stabilized to $\mathbf{t}$. Therefore, if $v(A \vee B) = \mathbf{t}$, then $v(A) = \mathbf{t}$ or $v(B) = \mathbf{t}$.

The equivalence of $\sim(A \,\&\, B)$ and $\sim A \vee \sim B$ secures the other part of the definition of K-valuation. □

I will note that from the assumption that neither A nor B is stably $\mathbf{t}$ in a valuation, pairs of sequences, as in the proof, can be constructed arbitrarily far into ω_1. So $A \vee B$ will receive a non-$\mathbf{t}$ value arbitrarily far out.

It follows from the preceding theorem that ultimate value and stable value coincide for $V(\omega_1)$. The ordinals that we are looking for are at least uncountably long. To diagnose the problem with countable sequences properly, I need one more definition from set theory.

Definition 35. *The* cofinality of α, $cf(\alpha)$, *is the least ordinal that can be mapped unbounded into* α.[48]

For all ordinals α, $cf(\alpha)$ is a cardinal no greater than α and, in particular, $cf(\omega_1) = \omega_1$. The proof will work for any ordinal α such that the $cf(\alpha) > \omega$. It is important for the argument that $cf(\omega_1) > \omega$. We can construct a valuation for uncountable sequences that is not a K-valuation if the length of the sequences has cofinality ω, such as ω_ω. Let S be the sequence countable sequence $\langle 0, \omega, \omega_1, \omega_2, \ldots, \omega_n, \ldots \rangle$, so $sup(S) = \omega_\omega$. Let the value of A and B be $\mathbf{t}$ from 0 to ω. From $\omega + 1$, let B be $\mathbf{f}$ and keep A at $\mathbf{t}$. At the successor of each limit in the sequence switch the values of A and B until the successor of the next limit in S. This assignment continues through S. The result is a valuation that assigns stably $\mathbf{t}$ to $A \vee B$ but assigns stable values to neither A nor B, as desired.

The proof of theorem 32 is heavy-handed. The propositional fragment cannot define many sequences of values. Since all countable limit ordinals have cofinality ω, it should be possible to carry out a construction similar to the previous one for all countable limit ordinals. This bars us from obtaining an ordinal smaller than ω_1 for which all valuations are K-valuations.

It appears that the appropriate framework for investigating Field's logic is the one based on $V(\omega_1)$ or $V(\lambda)$, where $cf(\lambda) > \omega$. $Values(\omega_1)$ contains uncountably many values, which

[48] This definition is based on Kunen (1983, 32). See Kunen (1983, 32-35) for more properties of cofinality.

are partially ordered.

Now that we have found a candidate space as the semantic values for the study of Field's conditional, there is a question about what it tells us about Field's conditional. The set $V(\omega_1)$ validates all of the laws and rules involving just propositional connectives listed in Field (2008, Ch. 17.4). Additionally, it invalidates all of the invalid sentences and rules listed in that section. Beyond that, here are a few additional rules that are valid on this semantics that are not listed anywhere by Field.

- $A, B \to C, B \to \sim C \models \sim(A \to B)$
- $A \models (A \,\&\, B) \leftrightarrow B$
- $\sim(C \to A \to .C \to B) \models \sim(A \to B)$
- $\sim(C \to A \to .C \to B) \models (C \to A) \,\&\, \sim(C \to B)$
- If $\models (A \,\&\, B) \to C$, then $A \models B \to C$

I will omit the proofs since they involve lots of tedious case checking.

There are many further questions for the framework. The most interesting are the completeness questions. Is there a complete axiomatization of validity based on $V(\omega_1)$? Is there a complete axiomatization of validity based on $V(\lambda)$, for all λ with $cf(\lambda) > \omega$? Is there a complete axiomatization of validity based on $V(\lambda)$ for non-successor limit ordinals λ? We do not have the answers to any of these questions.

6.5 CONCLUSIONS

In this chapter I have presented Field's theory of truth. I argued that it is inadequate by my criteria and that it does not cohere with Field's philosophical views. One of the problems with the theory is that its conditional lacks an axiomatization. The way the conditional is bound up with the paradoxical behavior of the truth predicate presents some obstacles to studying the conditional in isolation. I proposed above a framework for studying the conditional without the added complications of Field's canonical models. Most of the interesting questions about this logic remain open.

7.0 CONCLUDING THOUGHTS

We have covered a lot of ground, with many argumentative threads and lengthy excursions into formal work. I would like to close by summarizing what has been accomplished, along the way indicating future directions in which to take this work.

The broad problem of Semantic Closure features in criticisms of many theories of truth. Different versions of these criticisms impose different requirements on theories. I have broken the problem into six parts and examined major arguments for requirements stemming from each part. From the standpoint of the descriptive project, many of these arguments fail to get a grip, although there are some that are still forceful and provide us with substantive requirements. I will reiterate my positive conclusions, rather than focus on the negative ones, except for the following comment.

I argued that from the starting point of the descriptive project, many of the requirements imposed by the broad problem of Semantic Closure can be resisted. The revision theory, a version of which I developed and defended in chapters 3 and 4, has been criticized along many of the lines I argued against. I think that approaches besides the revision theory should be amenable to the negative conclusions I reached. In particular, contextualists, such as Glanzberg and Simmons, could use defenses along the lines I developed. We can now proceed to the positive conclusions.

I argued that Extensibility ($\forall\exists$) was true (§2.1). Using this, I argued that an adequate theory of truth should satisfy Logic Neutrality. This immediately provides a criticism of some fixed-point theories, since they cannot be used with certain logical resources.

I argued that an adequate theory of truth should satisfy closure conditions including Syntactic Closure (§2.2). The focus in the theory of truth is on languages that contain syntactic resources that systematically name or designate the sentences of the language.

Rich languages support natural strengthenings of this, such as including function symbols for manipulating sentence names. In addition, an adequate theory of truth should satisfy generalized semantic closure, which requires validating all the adequacy conditions on the semantic concepts. This requires, in particular, that the theory validate the T-sentences.

I distinguished two sorts of semantic predicates, semantic value and diagnostic (§2.3). Theories of truth all contain at least one semantic value predicate, and, if there are appropriate syntactic resources in a language, there will be a second semantic value predicate, "is false." I argued that a theory of truth that is adequate for the descriptive project should provide resources for diagnostically classifying sentences. A prime example is diagnosing the ways in which the defectiveness of the liar differs from that of the truth-teller. Many theories of truth provide robust tools for this and other classifications. Some approaches, such as that of Horsten and Field, take on philosophical commitments that preclude the use of such diagnostic tools, so they are subject to criticism on this basis.

I went on to argue that for theories of truth to be adequate, it is a necessary condition that they validate the semantical laws, assuming that there are appropriate syntactic resources in the language (§2.5). This condition is rooted in the idea that the theory of truth should be connected to the overall semantic theory for a language.

The positive conclusions I reached point towards future work. There are philosophical and technical issues to investigate with respect to logic neutrality, chief among which is settling its extent. In addition, there are questions concerning the diagnostic predicates to investigate. These include representing diagnostic notions in an object language with truth, broadening the scope to include new diagnostic categories, such as Curry-like, and cross-theoretic comparisons. For example, in a given model, we can define the class of intrinsically true sentences of the strong Kleene fixed-point theory. Can the revision theory tell us anything informative about them? Recently work has been done to clarify different notions of dependence and grounding.[1] Can we define these notions in the revision theory? What are the limits of the possibilities of combining diagnostic categories and truth? Answers to these questions would increase our understanding of semantic notions.

Reflection on the positive conclusions highlights the roles of conditionals for theories

[1] See Yablo (1982), Leitgeb (2005), and Meadows (2013).

of truth. In particular, both the T-sentences and the semantical laws essentially use a biconditional connective. To satisfy the descriptive project, a theory of truth needs to be able to reconstruct arguments, many of which will use conditionals in an important way. This raises the question of what a good conditional for a theory of truth is.

I investigated one answer to this question in chapter 3, the step conditionals. The T-sentences formulated using the step biconditional are valid, which fixes the main problem with the revision theory. More generally, if $\mathscr{D}$ is a set of definitions, then for each definitional equivalence in $\mathscr{D}$, there is a corresponding valid step biconditional. This is the basis for the claim that the step biconditional *reflects* definitional equivalence. There is more to the story of the step conditionals, since there are valid step conditionals and biconditionals to which no definitional equivalences correspond. On the philosophical view I defend, the appropriate relation between the step biconditional and definitional equivalence is that of reflecting, rather than, say, the stronger relation of expressing. In later work, I hope to further develop and sharpen the philosophical picture of step conditionals presented here.

The step conditionals each have a distinctive logic different from that of the classical material conditional (§3.2). This led to the worry that they are inadequate as conditionals and that the revision theory with their addition is no better off than without (§3.4.3). To respond, I distinguished two roles for conditionals in theories of truth, the ordinary reasoning role and the truth-related role. Neither the material conditional alone nor the step conditionals alone fill both roles, but together they do so adequately.

I went on to investigate the unary operator, $\Box$, that can be used to define the step conditionals using the classical material conditional. I proved soundness (§4.2) and completeness (§4.3) claims for a modification of C_0 with respect to to S_0 validity. The completeness claim also holds for $S^{\#}$ validity as long as the set of definitions under consideration is finite.

The box has a simple modal logic, which I called RT. I formulated a sequent system for the propositional logic and a hypersequent system for the quantified logic with Barcan and converse Barcan formulas (§5.1). I proved Solovay-style completeness results for the propositional and quantified forms of the logic (§5.2). Given the connections between the modal logic and circular definitions in the revision theory, it seems appropriate to say that RT is the modal logic of revision.

Finally, I looked at Hartry Field's recently proposed theory of truth in order to compare it with the modified revision theory and see how it fares according to my criteria. Field's theory adds a new conditional to the basic strong Kleene fixed-point theory to strengthen the logic. Field's theory does not satisfy all the requirements for which I argued in earlier chapters. In particular, it does not satisfy logic neutrality. I went on to argue that Field's theory does not support his philosophical views. Artifacts of Field's formal construction undermine his deflationism (§6.3.2), and Field's theory does not support his claims about truth-preservation (§6.3.3). In addition, the logic of Field's conditional is opaque. Its logical behavior depends on features of Field's models that build in a great deal of complexity at the outset. In order to better understand the logic, I proposed a framework for studying Field's conditional apart from his models and investigated some of the features of this framework (§6.4). My framework shed some light on the logic of Field's conditional, but questions of axiomatization were left open. An axiomatization of the logic would be helpful for evaluating Field's theory.

Field's theory turns out to satisfy neither my criteria nor even his own. The modified revision theory that I defend satisfies my criteria, and compares favorably to Field's theory. The modified revision theory also provides a general theory of circular definitions, not just a theory of truth. The modified revision theory looks strong on my analysis of the broad problem of Semantic Closure.

8.0 BIBLIOGRAPHY

Anderson, A. and Belnap, N. (1975). *Entailment: The Logic of Relevance and Necessity*, volume 1. Princeton University Press.

Avron, A. (1987). A constructive analysis of RM. *Journal of Symbolic Logic*, 52(4):939–951.

Beall, J. (2009). *Spandrels of Truth*. Oxford University Press.

Beall, J. and Murzi, J. (2013). Two flavors of Curry's paradox. *Journal of Philosophy*, 110(3):143–165.

Belnap, N. (1962). Tonk, plonk and plink. *Analysis*, 22(6):130–134.

Belnap, N. (1982a). Display logic. *Journal of Philosophical Logic*, 11(4):375–417.

Belnap, N. (1982b). Gupta's rule of revision theory of truth. *Journal of Philosophical Logic*, 11(1):103–116.

Belnap, N. (2009). Notes on the science of logic. Unpublished.

Belnap, N. and Müller, T. (2013). CIFOL: Case-intensional first order logic. *Journal of Philosophical Logic*, Forthcoming.

Blackburn, P., de Rijke, M., and Venema, Y. (2002). *Modal Logic*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

Boolos, G. (1993). *The Logic of Provability*. Cambridge University Press.

Brady, R. (2006). *Universal logic*. CSLI Publications.

Burge, T. (1979). Semantical paradox. *Journal of Philosophy*, 76(4):168–198.

Caret, C. and Cotnoir, A. J. (2008). True, False, Paranormal and 'Designated'?: A Reply to Jenkins. *Analysis*, 68(299):238–244.

Chihara, C. (1979). The semantic paradoxes: A diagnostic investigation. *Philosophical Review*, 88(4):590–618.

Cobreros, P., Egre, P., van Rooij, R., and Ripley, D. (2013). Reaching transparent truth. *Mind*, Forthcoming.

DeVidi, D. and Solomon, G. (1999). Tarski on "essentially richer"metalanguages. *Journal of Philosophical Logic*, 28(1):1–28.

Dunn, J. M. (1995). Positive modal logic. *Studia Logica*, 55(2):301–317.

Feferman, S. (1984). Toward useful type-free theories. I. *Journal of Symbolic Logic*, 49(1):75–111.

Field, H. (1999). Deflating the conservativeness argument. *Journal of Philosophy*, 96(10):533–540.

Field, H. (2003). A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic*, 32:139–177.

Field, H. (2004). The semantic paradoxes and the paradoxes of vagueness. In Beall, J., editor, *Liars and Heaps: New Essays on Paradox*, pages 262–311. Oxford University Press.

Field, H. (2006a). Maudlin's *Truth and Paradox*. *Philosophy and Phenomenological Research*, 73(3):713–720.

Field, H. (2006b). Truth and the unprovability of consistency. *Mind*, 115(459):567–605.

Field, H. (2008). *Saving Truth from Paradox*. Oxford.

Fitch, F. (1964). Universal metalanguages for philosophy. *Review of Metaphysics*, 17:396–402.

Fitch, F. B. (1946). Self-reference in philosophy. *Mind*, 55(217):64–73.

Glanzberg, M. (2004). A Contextual-Hierarchical Approach to Truth and the Liar Paradox. *Journal of Philosophical Logic*, 33(1):27–88.

Grover, D. L. (1992). *Prosentential Theory of Truth*. Princeton University Press.

Grover, D. L., Kamp, J. L., and Belnap, N. D. (1975). A prosentential theory of truth. *Philosophical Studies*, 27(1):73–125.

Gupta, A. (1978). Modal logic and truth. *Journal of Philosophical Logic*, 7(1).

Gupta, A. (1982). Truth and paradox. *Journal of Philosophical Logic*, 11(1):1–60.

Gupta, A. (1984). Truth and paradox. In Martin, R. L., editor, *Recent Essays on Truth and the Liar Paradox*, pages 175–236. Oxford University Press.

Gupta, A. (1997). Definition and revision: A response to McGee and Martin. *Philosophical Issues*, 8:419–443.

Gupta, A. (2006). Finite circular definitions. In Bolander, T., Hendricks, V. F., and Andersen, S. A., editors, *Self-Reference*, pages 79–93. CSLI Publications.

Gupta, A. and Belnap, N. (1993). *The Revision Theory of Truth*. MIT Press.

Gupta, A. and Martin, R. L. (1984). A fixed point theorem for the weak Kleene valuation scheme. *Journal of Philosophical Logic*, 13(2):131–135.

Halbach, V. (2011). *Axiomatic Theories of Truth*. Cambridge University Press.

Halbach, V. and Horsten, L. (2006). Axiomatizing Kripke's Theory of Truth. *Journal of Symbolic Logic*, 71:677–712.

Herzberger, H. G. (1982). Notes on naive semantics. *Journal of Philosophical Logic*, 11(1):61–102.

Horsten, L. (2009). Levity. *Mind*, 118(471):555–581.

Horsten, L. (2011). *The Tarskian Turn: Deflationism and Axiomatic Truth*. MIT Press.

Horsten, L., Leigh, G., Leitgeb, H., and Welch, P. D. (2012). Revision revisited. *Journal of Philosophical Logic*, forthcoming.

Hughes, G. E. and Cresswell, M. J. (1996). *A New Introduction to Modal Logic*. Routledge.

Jenkins, C. S. (2007). True, False, Paranormal and Designated: A Reply to Beall. *Analysis*, 67(1):80–83.

Jenkins, C. S. (2008). The importance of being designated: A comment on caret and cotnoir. *Analysis*, 68(3):244–247.

Kishida, K. (2010). *Generalized Topological Semantics for First-Order Modal Logic*. PhD thesis, University of Pittsburgh.

Koons, R. C. (1994). Review of the *Revision Theory of Truth*. *Notre Dame Journal of Formal Logic*, 35(4):606–631.

Kremer, M. (1986). *Logic and Truth*. PhD thesis, University of Pittsburgh.

Kremer, M. (1988). Kripke and the logic of truth. *Journal of Philosophical Logic*, 17:225–278.

Kremer, M. (2002). Intuitive consequences of the revision theory of truth. *Analysis*, 62(4):330–336.

Kremer, P. (1993). The Gupta-Belnap systems $S^{\#}$ and S^{*} are not axiomatisable. *Notre Dame Journal of Formal Logic*, 34(4):583–596.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72:690–716.

Kunen, K. (1983). *Set Theory: An Introduction to Independence Proofs*. North Holland.

Leitgeb, H. (2005). What truth depends on. *Journal of Philosophical Logic*, 34(2):155–192.

Lewis, D. (1975). Languages and language. In *Minnesota Studies in the Philosophy of Science*, volume 7, pages 3–35. University of Minnesota Press.

Martin, D. A. (1997). Revision and its rivals. *Philosophical Issues*, 8:407–418.

Martin, R. L. and Woodruff, P. W. (1975). On representing 'true-in-L' in L. *Philosophia*, 5(3):213–217.

Maudlin, T. (2004). *Truth and Paradox: Solving the Riddles*. Oxford University Press.

McGee, V. (1985). How truthlike can a predicate be? A negative result. *Journal of Philosophical Logic*, 14(4):399–410.

McGee, V. (1991). *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*. Cambridge.

McGee, V. (1997). Revision. *Philosophical Issues*, 8:387–406.

McGee, V. (2010). Field's logic of truth. *Philosophical Studies*, 147(3):421–432.

Meadows, T. (2013). Truth, dependence and supervaluation: Living with the ghost. *Journal of Philosophical Logic*, 42(2):221–240.

Patterson, D. (2007). Understanding the liar. In *Revenge of the Liar: New Essays on the Paradox*, pages 197–224. Oxford University Press.

Pottinger, G. (1983). Uniform cut-free formulations of T, S4 and S5 (abstract). *Journal of Symbolic Logic*, 48(3):900–901.

Priest, G. (1984). Semantic closure. *Studia Logica*, 43(1-2):117–129.

Priest, G. (2006). *In Contradiction: A Study of the Transconsistent*. Oxford University Press, 2nd edition.

Priest, G. (2010). Hopes fade for saving truth. *Philosophy*, 85(1):109–140.

Quine, W. (1976). Ways of paradox. In *Ways of Paradox*, pages 1–18. Harvard UP.

Ray, G. (2005). On the matter of essential richness. *Journal of Philosophical Logic*, 34(4):433–457.

Read, S. (2009). Plural signification and the liar paradox. *Philosophical Studies*, 145(3):363–375.

Read, S. (2010). Field's paradox and its medieval solution. *History and Philosophy of Logic*, 31(2):161–176.

Reinhardt, W. (1986). Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic*, 15:219–251.

Restall, G. (1992). A note on naive set theory in LP. *Notre Dame Journal of Formal Logic*, 33(3):422–432.

Restall, G. (1994). *On Logics Without Contraction*. PhD thesis, University of Queensland.

Restall, G. (2009). Truth values and proof theory. *Studia Logica*, 92(2):241–264.

Restall, G. (2012). A cut-free sequent system for two-dimensional modal logic, and why it matters. *Annals of Pure and Applied Logic*, 163(11):1611–1623.

Ripley, D. (2013). Revising up: Strengthening classical logic in the face of paradox. *Philosophers' Imprint*, 13(5):1–13.

Scharp, K. (2007). Aletheic vengeance. In *Revenge of the Liar: New Essays on the Paradox*, pages 272–319. Oxford University Press.

Shapiro, L. (2011). Expressibility and the liar's revenge. *Australasian Journal of Philosophy*, 89(2):297–314.

Shaw, J. (2013). Truth, paradox, and ineffable propositions. *Philosophy and Phenomenological Research*, 86:64–104.

Simmons, K. (1993). *Universality and the Liar: An Essay on Truth and the Diagonal Argument*. Cambridge University Press.

Tarski, A. (1969). Truth and proof. *Scientific American*, 220:63–77.

Tarski, A. (1983). The concept of truth in formalized languages. In Corcoran, J., editor, *Logic, Semantics, Metamathematics*, pages 152–278. Hackett.

Updike, E. (2010). *Paradise Regained: Fitch's Program of Basic Logic*. PhD thesis, UC Irvine.

Visser, A. (1989). Semantics and the liar paradox. In Gabbay, D. and Guethner, F., editors, *Handbook of Philosophical Logic*, volume 4, pages 617–706. Reidel.

Visser, A. (2004). Semantics and the liar paradox. In Gabbay, D. and Guethner, F., editors, *Handbook of Philosophical Logic*, volume 11, pages 149–240. Springer, 2nd edition.

Wansing, H. (1998). *Displaying Modal Logic*. Kluwer.

Weber, Z. (2010). Transfinite numbers in paraconsistent set theory. *Review of Symbolic Logic*, 3(1):71–92.

Welch, P. D. (2008). Ultimate truth vis-à-vis stable truth. *Review of Symbolic Logic*, 1(1):126–142.

Yablo, S. (1982). Grounding, dependence, and paradox. *Journal of Philosophical Logic*, 11(1):117–137.

Yablo, S. (2003). New grounds for naive truth theory. In Beall, editor, *Liars and Heaps: New Essays on Paradox*, pages 312–330. Oxford University Press.

Yaqūb, A. M. (1993). *The Liar Speaks the Truth: A Defense of the Revision Theory of Truth*. Oxford University Press.

Zardini, E. (2011). Truth without contra(di)ction. *The Review of Symbolic Logic*, 4(04):498–535.