Revenge of the Liar

New Essays on the Paradox

edited by

JC BEALL

Revenge of the Liar

This page intentionally left blank

Revenge of the Liar

New Essays on the Paradox

EDITED BY

JC Beall

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York

First published 2007

British Library Cataloguing in Publication Data
Data available
Library of Congress Cataloging in Publication Data
Data available
Typeset by Laserwords Private Limited, Chennai, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd, King's Lynn, Norfolk
ISBN 978–0–19–923390–8
ISBN 978–0–19–923391–5
10 9 8 7 6 5 4 3 2 1

CONTENTS

NOTES ON CONTRIBUTORS

JC Beall, Professor of Philosophy, University of Connecticut, and Arché Associate Research Fellow, University of St Andrews. In addition to articles and edited volumes on truth, paradox, and related issues, Beall is the author of *Logical Pluralism* with Greg Restall and the textbook *Possibilities and Paradox: An Introduction to Modal and Many-Valued Logic* with Bas C. van Fraassen. He is currently finishing a monograph on transparent truth and paradox (forthcoming with Oxford University Press). When not doing philosophy, Beall enjoys walking in the woods (and trying not to do philosophy), listening to music, and home schooling his cats.

Roy T. Cook, Visiting Assistant Professor, Villanova University, and Arché Associate Research Fellow, University of St Andrews. Cook's main research interests are philosophy of mathematics, philosophical logic, mathematical logic, and the philosophy of language. He has published papers on these topics in *Mind, The Monist, Journal of Philosophical Logic, Journal of Symbolic Logic, Notre Dame Journal of Formal Logic, Dialectica*, and *Analysis*, among other places. He is also editor of *The Arché Papers on the Mathematics of Abstraction*. Cook is currently working on a comprehensive dictionary of philosophical logic. When not thinking about logic and mathematics, Cook enjoys building sculptures and mosaics out of LEGO© bricks, and attempting to keep his three cats from disassembling them.

Matti Eklund, Assistant Professor, Sage School of Philosophy, Cornell University. Research interests: metaphysics, philosophy of language, and philosophy of logic. His publications include 'Inconsistent Languages' (*Philosophy and Phenomenological Research* 2002), 'What vagueness consists in' (*Philosophical Studies* 2005) and 'Carnap and ontological pluralism' (forthcoming in D. Chalmers, D. Manley, and R. Wasserman (eds.), *Metametaphysics*, Oxford University Press). His current research focuses on the implications of the liar and sorites paradoxes, and on various issues in metaontology.

Hartry Field, Silver Professor of Philosophy, New York University. Field is the author of *Science Without Numbers* (Blackwell 1980), which won the Lakatos Prize, of *Realism, Mathematics and Modality* (Blackwell 1989), and of *Truth and the Absence of Fact* (Oxford 2001). His current research interests include objectivity and indeterminacy, a priori knowledge, causation, and the semantic and set-theoretic paradoxes; he is currently finishing a monograph on truth and paradox (forthcoming with Oxford University

Press). Field hates writing either bibliographical or biographical blurbs, and so leaves that task to the editor.

Thomas Hofweber, Associate Professor of Philosophy, University of North Carolina at Chapel Hill. Hofweber mainly works in metaphysics, the philosophy of language and the philosophy of mathematics. Sample publications: *Inexpressible Properties and Propositions* (Oxford Studies in Metaphysics, vol. 2, 2006), 'Number determiners, numbers and arithmetic' (*Philosophical Review* 2005), 'Supervenience and object-dependent properties' (*Journal of Philosophy* 2005), 'A puzzle about ontology' (*Nous* 2005). At present he is working on a book on the domain and methods of metaphysics, in particular ontology.

Hannes Leitgeb, Reader in Mathematical Logic and Philosophy of Mathematics, University of Bristol. Leitgeb's chief research interests are in philosophical logic, epistemology, cognitive science, and philosophy of mathematics. He has published, among others, in *Journal of Philosophical Logic*, *Synthese*, *Analysis*, *Philosophia Mathematica*, *Erkenntnis, Journal of Logic, Language and Information, Studia Logica*, *Notre Dame Journal of Formal Logic*, *Topoi, Logique et Analyse, Artificial Intelligence*. In 2004 he published *Inference on the Low Level: An Investigation into Deduction*, *Nonmonotonic Reasoning*, *and the Philosophy of Cognition* in the Kluwer/Springer Applied Logic series. At present, Leitgeb is trying to resurrect Carnap's Logical Structure of the World. Apart from philosophy he likes music and marathons.

Tim Maudlin, Professor of Philosophy, Rutgers University. Professor Maudlin's primary interest is in the nature of reality. Recent books include *Truth and Paradox* and *The Metaphysics Within Physics*. He is inordinately fond of Belgian waffles with whipped cream.

Douglas Patterson, Associate Professor of Philosophy, Kansas State University. Patterson works in the philosophy of language, philosophical logic, and related areas in metaphysics, epistemology, and the philosophy of mind. Patterson is the author of a number of articles on truth and related topics, including 'Theories of truth and convention T', *Philosophers' Imprint* 2:5 and 'Tarski, the liar and inconsistent languages', *The Monist* 89:1. He is the editor of a collection of essays on Tarski, forthcoming with Oxford University Press, and of *Inquiry* 50:6 on inconsistency theories of semantic paradox. In his spare time Patterson enjoys hiking and travel, and he was the 1986 Minnesota Junior Men's Biathlon champion.

Graham Priest, Boyce Gibson Professor of Philosophy, University of Melbourne, and Arché Professorial Fellow, University of St Andrews. Priest's chief research interests are logic and related areas, including metaphysics and the history of philosophy (East and West). He has published in nearly all the major philosophy journals. Recent books include, *Towards Non-Being, Doubt Truth to be a Liar*, and the second edition of *In Contradiction* (all with Oxford University Press). He is currently finishing a second

volume of his *Introduction to Non-Classical Logic*. When not doing philosophy, Priest enjoys doing philosophy.

Agustín Rayo, Associate Professor of Philosophy, MIT, and Arché Associate Research Fellow, University of St Andrews. Rayo's chief research interests are in philosophical logic and philosophy of language. Rayo has published in such areas in various journals, including *Nous* and the *Journal of Symbolic Logic*. He recently edited *Absolute Generality* with Gabriel Uzquiano. He is currently engaged in a research project on content. When he isn't doing philosophy, Rayo enjoys scuba diving and opera.

Stephen Read, Reader in History and Philosophy of Logic, University of St Andrews. His chief research interests are in philosophical logic, medieval logic, and metaphysics, in particular, the notion of logical consequence; and extend from medieval theories in the philosophy of language, mind, and logic, to the more modern concerns of relevance logic and the philosophy of logic. He has published in such areas in *History and Philosophy of Logic, Journal of Philosophical Logic, Mind, Philosophy, Vivarium*, and others. His books include *Relevant Logic* (1987) and *Thinking about Logic* (1995). He is currently working on a critical edition and English translation of Thomas Bradwardine's *Insolubilia*. When not doing philosophy, he enjoys opera, cycling, and playing the clavichord.

Greg Restall, Associate Professor of Philosophy, University of Melbourne. Restall's research interests are in logic, metaphysics, and related fields. Recent books include *Logic* (a textbook) and *Logical Pluralism,* with co-author JC Beall, and he is currently working on the connections between proof theory and meaning. Restall enjoys looking after his young son.

Kevin Scharp, Assistant Professor of Philosophy, The Ohio State University. Scharp's main areas of research are philosophy of language and philosophical logic. He has published papers in *British Journal for the History of Philosophy, International Journal of Philosophical Studies*, and *Inquiry*. Scharp is editor (along with Robert Brandom) of the forthcoming volume of Wilfrid Sellars' essays entitled *In the Space of Reasons* from Harvard University Press. Scharp is currently working on a book on truth and the liar paradox. When not doing philosophy, he loves backpacking, cooking, and listening to music.

Stewart Shapiro, O'Donnell Professor of Philosophy, The Ohio State University, and Arché Professorial Fellow, University of St Andrews. He specializes in philosophy of mathematics, philosophical logic, and philosophy of language, notably vagueness. His most recent book is *Vagueness in Context* (Oxford: Oxford University Press, 2007) and he is the editor of the *Oxford Handbook of the Philosophy of Mathematics and Logic*. He still jogs (if you can still call it that), and likes the Grateful Dead and Incredible String Band.

Keith Simmons, Professor of Philosophy, University of North Carolina at Chapel Hill. He has research interests in logic, philosophy of logic, and philosophy of language. He

is the author of *Universality and the Liar* (Cambridge University Press) and, with Simon Blackburn, the editor of *Truth* (in the Oxford Readings in Philosophy series). He is currently at work on a monograph about the paradoxes, and, with Dorit Bar-On, a monograph about truth.

P. D. Welch, Professor of Pure Mathematics, University of Bristol. He has research interests in set theory, models of computation, theories of truth, and possible worlds semantics. His work on these and related topics have appeared in many of the leading logic and mathematics journals. Welch is currently working with the The Luxemburger Zirkel on a major research project entitled *Logical Methods in Epistemology, Semantics, and Philosophy of Mathematics*.

1

Prolegomenon to Future Revenge

JC Beall

This chapter attempts to lay out some background to the target phenomenon: the Liar and its revenge. The phenomenon is too big, and the literature (much) too vast, to give anything like a historical summary, or even an uncontroversial sketch of the geography. Accordingly, my aim is simply to lay out a few background ideas, in addition to briefly summarizing the contributed essays. I also try to avoid overlap with the individual chapters' rehearsals of revenge (including standard references to historical theses, like 'semantic self-sufficiency'), as the chapters do a nice job covering such material. Finally, because some of the ideas are presupposed by many of the chapters in this volume, a gentle sketch of Kripke's 'fixed point' approach to truth is given in an appendix.

1.1 Truth

Whatever else it may do, *truth* is often thought to play Capture and Release. Where $Tr(x)$ is our truth predicate, α a sentence, and $\ulcorner\alpha\urcorner$ a name of α, Capture and Release are as follows.

$$\textsc{Capture}\text{: } \alpha \Rightarrow Tr(\ulcorner\alpha\urcorner)$$

$$\textsc{Release}\text{: } Tr(\ulcorner\alpha\urcorner) \Rightarrow \alpha$$

When $\Rightarrow$ is a *conditional*, we have the *Conditional Form* of Capture and Release: namely, when conjoined, the familiar T-biconditionals. When $\Rightarrow$ is a *turnstile*, we have the *Rule Form* of Capture and Release, which indicates 'valid inference' (in some sense).

The names 'Capture' and 'Release' arise from the fact that *Tr(x) captures* the information in *x*, fully storing it for its eventual *release*. In practice, a familiar—if not *the*—role of *Tr(x)* is its release function: an assertion of $Tr(\ulcorner \alpha \urcorner)$ releases all of the information in α. This is useful for 'long generalizations' or 'blind generalizations' or the like, many of which generalizations would be practically impossible if we didn't enjoy a truth predicate that played (at least the rule form of) Capture and Release.[1] That truth plays Capture and Release in Conditional Form is plausible but controversial. That truth plays Capture and Release in at least Rule Form is less controversial, and will henceforth be assumed.[2]

1.2 The Liar

The Liar phenomenon involves sentences that imply their own falsity or, more generally, untruth. By way of example, consider the ticked sentence in §1.2 of this chapter.

√ The ticked sentence in §1.2 of 'Prolegomenon to future revenge' is not true.

Assume that the ticked sentence is true. Release, in turn, delivers that the ticked sentence is not true. Hence, the ticked sentence, given Release (an essential feature of truth), implies its own untruth.

Is the ticked sentence untrue? Capture gives reason for pause: that the ticked sentence is not true implies, via Capture, that the ticked sentence is true!

The question is: what shall we say about the 'semantic status' of the ticked sentence? Answering this question invites the Liar's revenge.

1.3 The Liar's Revenge

On one hand, the *revenge phenomenon*—the Liar's revenge—is not so much a distinct phenomenon from the Liar as it is a witness to both the difficulty and ubiquity of Liars.

[1] Depending on the language, Rule Capture and Release is insufficient for a fully *transparent* truth predicate, one such that *Tr(x)* and *x* are intersubstitutable in all (non-opaque) contexts, for *all* sentences *x* in the language. By my lights, the Liar phenomenon is at its most difficult incarnation when truth is fully transparent, since any distinction between, for example, 'Excluded Middle' and 'Bivalence' collapses. But I will set this aside here. See Appendix for one approach to transparent truth, and see Field's chapter (Chapter 4) for another, stronger approach, as well as relevant discussion.

[2] A variety of theories reject even Rule Form of Capture and Release, but it will be assumed throughout this 'introduction'. One of the better known examples of rejecting Rule Capture is so-called Kripke–Feferman [3]. (See too [16].)

On the other hand, 'revenge' is often launched as an *objection* to an account of truth (or a response to Liars). Without intending a stark distinction, I will discuss *revenge qua Liar phenomenon* and *revenge qua objection* separately, with most of the discussion on the latter but all of the discussion brief.

1.3.1 The revenge phenomenon

The revenge phenomenon arises at the point of classifying Liars. Consider, again, the ticked sentence. As above, classifying the ticked sentence as *true* results in inconsistency; Release delivers that the ticked sentence is also not true. Likewise, classifying the ticked sentence as *not true* results in inconsistency; Capture delivers that the ticked sentence is also true. How, then, shall we classify the ticked sentence?

A natural suggestion is that the ticked sentence is *neither true nor false*. The trouble with this suggestion—even apart from logical issues involving negation—is the apparent connection between *being neither true nor false* and *being not true*.[3] In particular, presumably, we have

NTF−NT. $\quad \neg Tr(\ulcorner \alpha \urcorner) \wedge \neg Tr(\ulcorner \neg \alpha \urcorner) \Rightarrow \neg Tr(\ulcorner \alpha \urcorner)$

Again, $\Rightarrow$ may be a conditional or a turnstile. Either way, the problem at hand is plain. Assume, as per the current suggestion, that the ticked sentence is 'neither true nor false'. By NTF−NT, we immediately get that the ticked sentence is not true. But, now, we're back to inconsistency, as Capture, in turn, delivers that the ticked sentence is (also) true. So, while natural, the suggestion that the ticked sentence is *neither true nor false* is not a promising proposal.[4]

Quick reflection leads to a general lesson: whatever category one devises for the ticked sentence, it had better not imply untruth. For example, suppose that one introduces the category *bugger* for Liars. On this proposal, the ticked sentence is a *bugger*. Whatever else *being a bugger* might involve, we had better not have BUG if we're to avoid inconsistency.

BUG. $\quad Bugger(\ulcorner \alpha \urcorner) \Rightarrow \neg Tr(\ulcorner \alpha \urcorner)$

The trouble with BUG is exactly the trouble with NTF−NT. On the current proposal, the ticked sentence is a bugger, in which case, via BUG, it is not true. Capture, as before, delivers that the ticked sentence is (also) true, and inconsistency remains.

[3] Throughout, I will assume that *falsity* is truth of negation—i.e., that α is false just if $\neg\alpha$ is true. (This is a standard line, but it might be challenged. Fortunately, in the present context, nothing substantive turns on the issue.)

[4] I should note that my presentation simplifies matters a great deal. One might postulate a different negation at work in (wide-scope positions in) NTF−NT, thereby complicating matters. Moreover, one might—perhaps with some philosophical motivation—reject NTF−NT altogether. And there are other options, as will be evident in various chapters of this volume.

One lesson, then, is that our classification of the ticked sentence cannot consistently deliver its untruth. With the lesson in mind, suppose that we classify the ticked sentence as a *bugger* but, whatever else 'bugger' might mean, we reject BUG (in both Rule and Conditional Forms). Notwithstanding further details on 'buggerhood', this course yields the promise of consistently classifying the ticked sentence (and its negation): it is a bugger.

The *revenge phenomenon re-emerges*. Having, as we're assuming, consistently classified the ticked sentence as a 'bugger', *other Liars* emerge to thwart our aims at consistently (and completely) classifying Liars. By way of example, consider the starred sentence.

> ★ The starred sentence in §1.3.1 of 'Prolegomenon to future revenge' is either not true or a bugger.

Assume that the starred sentence is true. Release delivers that the starred sentence is not true or a bugger, and hence true *and* either not true or a bugger. Similarly, that the starred sentence is either not true or a bugger implies, via Capture, that it is true—and, hence, true *and* either not true or a bugger. Accordingly, given normal conjunction and disjunction behavior, if we have it that the starred sentence is either true, not true, or a bugger, we have either inconsistency (viz., true and not true) or some true buggers. While the latter option, without BUG (or similar principles), might afford a consistent theory, it is prima facie objectionable if, whatever else 'bugger' might mean, the buggers are thought to be somehow 'defective', sentences that ought to be rejected.[5]

The revenge phenomenon, at least in one relevant sense, is as above: it is not so much a separate phenomenon from the Liar as it is what makes the Liar phenomenon challenging. The Liar's revenge is reflected in the apparent hyrdra-like appearance of Liars: once you've dealt with one Liar, another one emerges. In short, if one manages to consistently classify a Liar as a *such-n-so*, another Liar emerges—e.g. a sentence that says of itself only that it's not true or a such-n-so. Dramatically and very generally put, Liars attempt to wreak inconsistency in one's language. If the Liar can't have what she wants, she'll enlist 'strengthened' relatives to frustrate your wants, in particular, your expressive wants. As it is sometimes put, Liars force—or try to force—you to choose between either inconsistently expressing what you want to express or not expressing what you want to express.

A *quietist* advises that we give up on our aim to classify Liars; there are Liars in the language, but there is no 'semantic category' in which the ticked sentence may truly be said to reside. Accordingly, whereof one cannot truly classify, thereof one must—or, in any event, might as well—be silent. The virtue of such an approach

[5] Whether 'buggers' should be conceived as defective (in some sense) is an open issue. Field's chapter (see Chapter 4) is relevant to this issue. See too [1].

is that it avoids revenge (since it doesn't engage); the salient defect is that it offers no clear account of truth or the paradoxes at all. Against such a 'proposal', little can be said, and so won't.

Another—so-called *dialetheic*—option is to *accept* the apparent inconsistency engendered by Liars. Provided that our logic tolerates such inconsistency—and part of the proposed lesson of the Liar is that our logic *does* tolerate such inconsistency—there's no obvious problem. What Liars teach us, on the dialetheic view, is that truth is inconsistent—that some true sentences have true negations. Whether such a position avoids the Liar's revenge is an open question.[6]

And there are (many) other options, as subsequent chapters reflect. What is uncontroversial is that the revenge phenomenon has fueled, and continues to fuel, work on the Liar phenomenon. This is not surprising, at least if, as suggested, the revenge phenomenon just is the Liar phenomenon—indeed, as above, a witness to the Liar's ubiquity.

1.3.2 Towards revenge qua objection

The literature on truth and paradox exhibits a familiar and ubiquitous pattern: each proposed 'account of truth' is followed by a charge of *revenge*, that the account can't accommodate such and so a notion (e.g. 'untruth', 'exclusively false', or whathaveyou) and, in that respect, is thereby inadequate. Indeed, were it not for alleged 'revenge' problems, many proposed theories of truth might be objection-free—or, at least, the number of known or cited objections would be greatly diminished.

Such 'revenge' charges, as said, are often launched as inadequacy objections against proposed accounts of truth. Unfortunately, there is some unclarity about the relevance of such charges, and, more to the point, unclarity with respect to the burden involved in successfully establishing the intended inadequacy result. Without aiming to resolve them, §1.4 briefly discusses some of the given issues involved in *revenge qua objection*. Before turning to §1.4, two background issues need to be briefly covered.[7]

1.3.2.1 Incoherent operators

By way of background, it is important to see that there are operators that cannot coherently exist if our language enjoys various features. Tarski's Theorem gives one

[6] That some sentences are true and false is one thing; however, the dialetheic position is rational only if at least some sentences are *just true*. The worry is whether the dialetheist can give an adequate account of 'just true' without the position exploding into triviality. Some of the chapters have discussion of this point. For a general discussion (and defense) of dialetheism, see [14, 15].

[7] I should warn that, from this point forward, my presentation may border on controversial.

concrete example of such a result,[8] but another example might be useful. In particular, suppose, as is plausible, that our language has features F1 and F2.

F1. There's a predicate *Tr*(x) that 'obeys' (unrestricted) Release and Capture in at least Rule Form.

F2. 'Reasoning by Cases' is valid: if α implies γ, and β implies γ then $\alpha \vee \beta$ implies γ, for all α, β, γ.

As such, the language, on pain of triviality, has no operator Φ such that both E1 and E2 hold.[9]

E1. $\vdash \alpha \vee \Phi\alpha$

E2. $\alpha, \Phi\alpha \vdash \bot$

Suppose that we do have such an operator. Consider a familiar construction, which will be guaranteed via diagonalization, self-reference or the like: a sentence λ that 'says' $\Phi Tr(\ulcorner\lambda\urcorner)$. From E1, we have

$$Tr(\ulcorner\lambda\urcorner) \vee \Phi Tr(\ulcorner\lambda\urcorner)$$

which yields two cases.

1. Case one:
   (a) $Tr(\ulcorner\lambda\urcorner)$
   (b) Release yields:[10] $\Phi Tr(\ulcorner\lambda\urcorner)$.
   (c) E2 yields: $\bot$
2. Case two:
   (a) $\Phi Tr(\ulcorner\lambda\urcorner)$
   (b) Capture yields: $Tr(\ulcorner\lambda\urcorner)$
   (c) E2 yields: $\bot$

The point, for present purposes, is modest but important: there are incoherent notions, notions that cannot coherently exist if our language enjoys various features. While modest, the point is something on which all parties can agree.

[8] Tarski's Theorem, in effect, is that (classical) arithmetical truth is not definable in (classical) arithmetic. For a user-friendly discussion of the theorem and its broader implications, see [20] and, more in-depth, [18]. For a user-friendly discussion of what Tarski's Theorem does *not* teach us, see [23], which is also highly relevant to 'revenge' issues, in general, and particularly relevant to Field's proposal (see Chapter 4).

[9] E1 might be thought of as an *exhaustion* principle, and E2 as *exclusion* or *explosion* principle. Throughout, $\bot$ is an 'explosive' sentence, one that implies all sentences.

[10] Intersubstitutability of Identicals is also involved here (and at the same place in Case two). This is usually assumed to be valid, but it, like so much in the area, has been challenged. See [17].

A principal question, at the heart of Liar studies, is this: what is our language like, given that it enjoys *such and so* features? More to the point: assuming that our language has a truth predicate that plays Capture and Release (in at least rule form), what are its other features? One might say that it fails to contain a fully *exhaustive* device, something that would yield E1, or fails to have any fully *explosive* device, something that would yield E2. One might, with various theorists, say that F2, in its given unrestricted form, fails for our language. One might say other things.

Whatever one says, one aims to give a clear, precise account of the matter—a clear, precise account of what our language is like, given that it has such and so features. This is normally done by way of a 'formal modeling'.

1.3.2.2 Models and reality

Like much in philosophical logic, constructing a formal account of truth is 'model building' in the ordinary 'paradigm' sense of 'model'. The point of such a model is to indicate how 'real truth' in our 'real language' can have the target (logical) features we take it to have—e.g. consistency (or, perhaps, inconsistency but non-triviality), Release and Capture features, perhaps full intersubstitutability of $Tr(\ulcorner\alpha\urcorner)$ and α. In that respect, formal accounts of truth are idealized models to be evaluated by their adequacy with respect to the 'real phenomena' they purport to model.[11]

Formal accounts (or theories) of truth aim only indirectly at being accounts of truth. What we're doing in giving such an account is two-fold.

1. We construct an artificial *model language*—one that's intended to serve as a heuristic, albeit idealized, model of our own 'real' language—and, in turn, give an account of how 'true' behaves in that language by constructing a precise account of *truth-in-that-language*.
2. We then claim that the behavior of 'true' in our language, at least in relevant, target respects, is like the behavior of the truth predicate in our model language.

By far the most dominant approach towards the first task—viz. constructing one's model language—employs a classical set theory. One reason for doing so is that classical set theory is familiar, well-understood, and generally taken to be consistent. A related reason is that, in using a classical set theory, one's formal account of truth

[11] Theories, like McGee's [13], that purport not to be 'descriptive' but, rather, 'revisionary' or 'normative', are not typically subject to 'revenge'-charges to the same extent that 'descriptive' theories are, and so are not the chief concern here. On the other hand, McGee aims to give a revisionary theory (not to be confused with *revision theory*) that aims to stay as close to the phenomena—our 'real language'—as possible. In that respect, 'revenge' objections might well arise.

can be more than merely a heuristic picture; it can also serve as a 'model' in the technical sense of *establishing consistency*.[12]

That a classical set theory is used in constructing our artificial language serves to emphasize the heuristic, idealized nature of the construction. We know that, due to paradoxical sentences, there's no truth predicate in (and for) our 'real language' if our real language is (fully) classical.[13] But the project, as above, is to show how we can have a truth predicate in our 'real language', despite such paradoxical sentences. And the project, as above, is usually—if not always—carried out in a classical set theory. Does this mean that the project, as typically carried out, is inexorably doomed? Not at all. Just as in physics, where idealization is highly illuminating despite its distance from the real mess, so too in philosophical logic: the classical construction is illuminating and useful, despite its notable idealization. But it is idealized, and, pending argument, on the surface only heuristic. That's the upshot of using classical set theory.

1.4 Comments on Revengers' Revenge

A quick glance at the Liar literature will indicate that 'revenge' is often invoked as a *problem* for a given theory of truth and paradox. For present purposes, a *revenger* is one who charges 'revenge' against some proposed account of truth. The principal issue of this section is the burden of revenge—the burden that revengers carry. The chapters in this volume will tell their own (and not necessarily compatible) story on this issue.

1.4.1 Too easy revenge

As above, in giving a formal theory of truth, one does not directly give a theory of *truth*; rather, one gives a theory of $\mathcal{L}_m$-truth, an account, for some formal 'model language' $\mathcal{L}_m$, of how $\mathcal{L}_m$'s truth predicate behaves, in particular, its logical behavior. By endorsing a formal theory of truth, one is endorsing that one's own truth predicate is relevantly like *that*, like the truth predicate in $\mathcal{L}_m$, at least with respect to various phenomena in question—for example, logical behavior.

[12] In paraconsistent contexts, the aim is basically the same, except that the target result is *non-triviality despite negation-inconsistency*. In the more dominant non-paraconsistent cases, the aim is also non-triviality, but that's ensured by consistency.

[13] The same applies, of course, if the truth predicate has an extension: the extension isn't really a classical set. Every classical set $\mathcal{S}$ is such that $x \in \mathcal{S} \vee x \notin \mathcal{S}$, which, given paradoxical sentences, results in inconsistency. (The point is independent of 'size' issues. Classical *proper classes* are likewise such that $x \in \mathcal{C} \vee x \notin \mathcal{C}$.) If $\mathcal{T}$ is the extension of $Tr(x)$ and $\mathcal{T}$ is a set, a sentence λ that 'says' $\ulcorner\lambda\urcorner \notin \mathcal{T}$ makes the point—assuming, as is plausible, suitable 'extension' versions of Capture and Release (e.g., $\alpha \Rightarrow \ulcorner\alpha\urcorner \in \mathcal{T}$, etc.).

Revenge qua objection—revenger's revenge—is an *adequacy objection*. Typically, the revenger charges that a given 'model language' is inadequate due to expressive limitation. Let $\mathcal{L}$ be our 'real language', English or some such natural language, and let $\mathcal{L}_m$ be our heuristic model language. Let '$\mathcal{L}_m$-truth' abbreviate 'the behavior of $\mathcal{L}_m$'s truth predicate'. In broadest terms, the situation is this: we want our (heuristic) $\mathcal{L}_m$, and in particular $\mathcal{L}_m$-truth, to illuminate relevant features of our own truth predicate, to explain how, despite paradoxical sentences, our truth predicate achieves the features we take it to have. Revenge purports to show that $\mathcal{L}_m$ achieves its target features in virtue of lacking expressive features that $\mathcal{L}$ itself (our real language) appears to enjoy. But if $\mathcal{L}_m$ enjoys the target features only in virtue of lacking relevant features that our real $\mathcal{L}$ enjoys, then $\mathcal{L}_m$ is an inadequate model: it fails to show how $\mathcal{L}$ itself achieves its target features (e.g. consistency). That, in a nutshell, is one common shape of revenge.

Consider a familiar and typical example, namely, Kripke's partial languages.[14] Let $\mathcal{L}_m$, our heuristic model language, be such a (fixed point) language constructed via the Strong Kleene scheme.[15] In constructing $\mathcal{L}_m$, we use—in our metalanguage—classical set theory, and we define *truth-in-*$\mathcal{L}_m$ (and similarly, false-in-$\mathcal{L}_m$), which notions are used to discuss $\mathcal{L}_m$-truth (the behavior of $\mathcal{L}_m$'s truth predicate). Moreover, we can prove—in our metalanguage—that, despite paradoxical sentences, a sentence $Tr(\ulcorner\alpha\urcorner)$ is true-in-$\mathcal{L}_m$ exactly if α is true-in-$\mathcal{L}_m$.

The familiar revenge charge is that $\mathcal{L}_m$, so understood, is not an adequate model; it fails to illuminate how our own truth predicate, despite paradoxical sentences, achieves consistency. In particular, the revenger's charge is that $\mathcal{L}_m$-truth achieves its consistency in virtue of $\mathcal{L}_m$'s expressive poverty: $\mathcal{L}_m$ cannot, on pain of inconsistency, express certain notions *that our real language can express*. Example: suppose that $\mathcal{L}_m$ contains a predicate $\varphi(x)$ that defines $\{\beta : \beta \text{ is not true-in-}\mathcal{L}_m\}$. And now, where λ says $\varphi(\ulcorner\lambda\urcorner)$, we can immediately prove—in the metalanguage—that λ is true-in-$\mathcal{L}_m$ iff $\varphi(\ulcorner\lambda\urcorner)$ is true-in-$\mathcal{L}_m$ iff λ is not true-in-$\mathcal{L}_m$. *Because—and only because—we have it (in our classical metalanguage) that* λ *is true-in-*$\mathcal{L}_m$ *or not*, we thereby have a contradiction: that λ is both true-in-$\mathcal{L}_m$ and not. But since we have it that truth-in-$\mathcal{L}_m$ is consistent (given consistency of classical set theory in which $\mathcal{L}_m$ is constructed), we conclude that $\mathcal{L}_m$ cannot express 'is not true-in-$\mathcal{L}_m$'.

The revenger's charge, then, amounts to this: that the Kripkean model language fails to be enough like our real language to explain at least one of the target phenomena, namely, truth's consistency. Our metalanguage is part of our 'real language', and we can define $\{\beta : \beta \text{ is not true-in-}\mathcal{L}_m\}$ in our metalanguage. As the Kripkean language

[14] See Appendix for a sketch of the Kripkean 'partial predicates' approach.

[15] The point applies to any of the given languages, but the K_3-construction (Strong Kleene) is probably most familiar.

cannot similarly define $\{\beta : \beta$ is not true-in-$\mathcal{L}_m\}$, the Kripkean model language is inadequate: it fails to illuminate truth's target features.

A revenger engages in 'too easy revenge' if the revenger only points to such a result without establishing its relevance.[16] The relevance of such a result is not obvious. After all, the given notion is a *classically constructed* notion; it is a 'model-dependent' notion—a notion that makes no sense apart from the given (classically constructed) models—defined entirely in a classical metalanguage. As such, the given notion, presumably, is not one of the target (model-independent, or 'absolute') notions in $\mathcal{L}$ that $\mathcal{L}_m$ is intended to model. The question, then, isn't whether there's some notion $\mathcal{X}$ (e.g., 'not true-in-$\mathcal{L}_m$') that is inexpressible—or, at least, not consistently expressible—in $\mathcal{L}_m$. The question is the relevance of such a result.

One might think that the relevance is plain. One might, for example, think that the semantics for $\mathcal{L}_m$ is intended to reflect the semantics of $\mathcal{L}$, our real language. Since the semantics of the former essentially involves, for example, *not true-in-*$\mathcal{L}_m$, the semantics of our real language must involve something similar—at least if $\mathcal{L}_m$ is an adequate model of our real language. But, now, since *not true-in-*$\mathcal{L}_m$ is (provably) inexpressible in $\mathcal{L}_m$, we should conclude that $\mathcal{L}_m$ is an inadequate model of our real language $\mathcal{L}$, since our real language can express its own semantic notions—i.e. the notions required for giving the semantics of our language.

Such an argument might serve to turn otherwise 'too easy revenge' into a plainly relevant and powerful objection; however, the argument itself relies on various assumptions that involve quite complex issues. For example, one conspicuous assumption is that the 'semantics' of $\mathcal{L}_m$ is intended to reflect the semantics of our real language $\mathcal{L}$. This needn't be the case. For example, suppose that one rejects that semantics—the semantics of our real language—is a matter of giving 'truth conditions' or otherwise involves some explanatory notion of truth. In the face of Liars (or other paradoxes), one still faces questions about one's truth predicate, and in particular its logical behavior. By way of answering such questions, one might proceed as above: construct a model language that purports to illuminate how one's real truth predicate enjoys its relevant features (e.g. Capture and Release) without collapsing from paradox. In constructing and, in turn, describing one's 'model language', one might give 'truth-conditional-like semantics' for the model language by giving 'truth-in-a-model conditions' for the language. If so, it is plain that the 'semantics' of the model language are not intended to reflect the 'real semantics' of one's real language; they may, in the end, be only tools used for illuminating the logic of our real language, versus illuminating the 'real semantics' of our real language. So, a critical assumption in the argument above—the argument towards the relevance of the given inexpressibility results—requires argument. Likewise, the assumption that our

[16] Thanks to Lionel Shapiro for very useful discussion on 'too easy revenge'.

real language can 'express its own semantic notions', the notions involved in 'giving the semantics' of our language, requires argument, argument that may turn, as with the first assumption, on difficult issues concerning the very 'nature of semantics'.[17]

A would-be revenger, involved in too easy revenge, would have it easy but too easy. What is (generally) easy is showing that some classically constructed notion is inexpressible—or, at least, not consistently expressible—in a (classically constructed) non-classical 'model language'. What is too easy is the thought that showing as much is sufficient to undermine the adequacy of the given model language. The hard part is clearly establishing the relevance of such inexpressibility results, that is, clearly substantiating the alleged inadequacy. The difficulty, as above, is that the alleged inadequacy often relies on very complicated issues—the 'nature of semantics', the role of given model-dependent notions, and more.

1.4.2 Revenger's recipes, in general

Towards clarifying the burden involved in launching revenger's revenge, it might be useful to lay out a few common recipes for revenge qua objection. For simplicity, let $\mathcal{L}_m$ be a given formal model language for $\mathcal{L}$, where $\mathcal{L}$ is our target, real language—the language features of which $\mathcal{L}_m$ is intended to illuminate. Let $M(\mathcal{L}_m)$ be the metalanguage for $\mathcal{L}_m$, and assume, as is typical, that $M(\mathcal{L}_m)$ is a fragment of $\mathcal{L}$. Then various (related) recipes for revenge run roughly as follows.[18]

Rv1. Recipe One.

- Find some semantic notion $\mathcal{X}$ that is used in $M(\mathcal{L}_m)$ to classify various $\mathcal{L}_m$-sentences (usually, paradoxical sentences).
- Show, in $M(\mathcal{L}_m)$, that $\mathcal{X}$ is not expressible in $\mathcal{L}_m$ lest $\mathcal{L}_m$ be inconsistent (or trivial).
- Conclude that $\mathcal{L}_m$ is explanatorily inadequate: it fails to explain how $\mathcal{L}$, with its semantic notion $\mathcal{X}$, enjoys consistency (or, more broadly, non-triviality).

Rv2. Recipe Two.

- Find some semantic notion $\mathcal{X}$ that, irrespective of whether it is explicitly used to classify $\mathcal{L}_m$-sentences, is expressible in $M(\mathcal{L}_m)$.
- Show, in $M(\mathcal{L}_m)$, that $\mathcal{X}$ is not expressible in $\mathcal{L}_m$ lest $\mathcal{L}_m$ be inconsistent (or trivial).
- Conclude that $\mathcal{L}_m$ is explanatorily inadequate: it fails to explain how $\mathcal{L}$, with its semantic notion $\mathcal{X}$, enjoys consistency (or, more broadly, non-triviality).

[17] Some of the chapters in this volume discuss this assumption, an assumption that often goes under the heading 'semantic self-sufficiency'. For arguments against such an assumption, see [7, 8].

[18] This is not in any way an exhaustive list of recipes!

Rv3. Recipe Three.

- Find some semantic notion $\mathcal{X}$ that is (allegedly) in $\mathcal{L}$. (Argue that $\mathcal{X}$ is in $\mathcal{L}$.)
- Argue that $\mathcal{X}$ is not expressible in $\mathcal{L}_m$ lest $\mathcal{L}_m$ be inconsistent (or trivial).
- Conclude that $\mathcal{L}_m$ is explanatorily inadequate: it fails to explain how $\mathcal{L}$, with its semantic notion $\mathcal{X}$, enjoys consistency (or, more broadly, non-triviality).

As above, a *revenger* is one who charges 'revenge' against a formal theory of truth, usually along one of the recipes above. The charge is that the model language fails to achieve its explanatory goals. In general, the revenger aims to show that there's some sentence in our real language $\mathcal{L}$ that *ought* to be expressible in $\mathcal{L}_m$ if $\mathcal{L}_m$ is to achieve explanatory adequacy. The question is: how ought one reply to revengers? The answer, of course, depends on the details of the given theories and the given charge of revenge. For present purposes, without going into such details, a few general remarks can be made.

The weight of Rv1 or Rv2 depends on the sort of $\mathcal{X}$ at issue. As in §1.3.2.2 and §1.4.1, if $\mathcal{X}$ is a classical, model-dependent notion constructed in a *proper fragment of* $\mathcal{L}$, then the charge of inadequacy is not easy to substantiate, even if the inexpressibility of $\mathcal{X}$ in $\mathcal{L}_m$ is easy to substantiate. In particular, if classical logic extends that of $\mathcal{L}_m$, then there is a clear sense in which you may 'properly' rely on a classical metalanguage in constructing $\mathcal{L}_m$ and, in particular, truth-in-$\mathcal{L}_m$. In familiar non-classical proposals, for example, you endorse that $\mathcal{L}$, the real, target language, is non-classical but enjoys classical logic as a (proper) extension, in which case, notwithstanding particular details, there is nothing prima-facie suspect about relying on an entirely classical fragment of $\mathcal{L}$ to construct your model language and, in particular, classical model-dependent $\mathcal{X}$s. But, then, in such a context, it is hardly surprising that $\mathcal{X}$, being an entirely classical notion, would bring about inconsistency or, worse, triviality, in the (classically constructed) *non-classical* $\mathcal{L}_m$.[19]

Because classical logic is typically an extension of the logic of $\mathcal{L}_m$, the point above is often sufficient to blunt, if not undermine, a revenger's charge, at least if the given recipe is Rv1 and Rv2. As in §1.4.1, the revenger must establish more than the unsurprising result that a (usually classically constructed) model-dependent $\mathcal{X}$ is expressible in $M(\mathcal{L}_m)$ but not in $\mathcal{L}_m$; she must show the relevance of such a result, which might well involve showing that some non-model-dependent notion—some relevant 'absolute' notion—is expressible in $\mathcal{L}$ but, on pain of inconsistency (or non-triviality), inexpressible in $\mathcal{L}_m$. And this task brings us to Rv3.

Recipe Rv3 is perhaps what most revengers are following. In this case, the idea is to locate a relevant *non*-model-dependent notion in $\mathcal{L}$ and show that $\mathcal{L}_m$ cannot, on pain of inconsistency (or triviality), express such a notion. The dialectic along these lines is delicate.

[19] For closely related discussion, see Field's chapter (Chapter 4) and also [4].

Suppose that Theorist proposes some formal theory of truth, and Revenger, following Rv3, adverts to some 'absolute' notion $\mathcal{X}$ that, allegedly, is expressible in $\mathcal{L}$. If, as I'm now assuming, Theorist neither explicitly nor implicitly invokes $\mathcal{X}$ for purposes of classifying sentences, then Revenger has a formidable task in front of her. In particular, without begging questions, Revenger must show that $\mathcal{X}$ really is an intelligible notion of $\mathcal{L}$.

For example, recall, from §1.3.2.1, the discussion of 'incoherent operators', and assume that Theorist proposes a theory that has features F1 and F2 (and, for simplicity, is otherwise normal with respect to extensional connectives). Let any operator Φ that satisfies E1 and E2 be an *EE device* (for 'exclusive and exhaustive'). Against a typical paracomplete (or paraconsistent) proposal,[20] an Rv3-type revenger might maintain that $\mathcal{L}$, our real language, enjoys an EE device. If the revenger is correct, then standard paracomplete and paraconsistent proposals are inadequate, to say the least. But the issue is: why think that the revenger is correct? Argument is required, but the situation is delicate. What makes the matter delicate is that many arguments are likely to beg the question at hand. After all, according to (for example) paracomplete and paraconsistent theorists, what the Liar teaches us is that, in short, *there is no EE device in our language*! Accordingly, the given revenger cannot simply point to normal evidence for such a device and take that to be sufficient, since such 'evidence' itself might beg questions against such proposals. On the other hand, if the given theorist cannot otherwise explain—or, perhaps, explain away—normal evidence for the (alleged) device, then the revenger may make progress. But the situation, as said, is delicate.

The difficulty in successfully launching Rv3 might be put, in short, as follows. Theorist advances $\mathcal{L}_m$ as a model of (relevant features of) $\mathcal{L}$, our real language. Rv3 Revenger alleges that $\mathcal{X}$ exists in $\mathcal{L}$, and shows that, on pain of triviality, $\mathcal{X}$ is inexpressible in $\mathcal{L}_m$. The difficulty in adjudicating the matter is that, as in §1.3.2.1, Theorist may reasonably conclude that $\mathcal{X}$ is incoherent (given the features of our language that Theorist advances). Of course, if Revenger could establish that we *need* to recognize $\mathcal{X}$, perhaps for some theoretical work or otherwise, then the debate might be settled; however, such arguments are not easy to come by.

The burden, of course, lies not only on the Rv3 Revenger; it also lies with the given theorist. For example, typical paracomplete and paraconsistent theorists must reject the intelligibility of any EE device in our language. Inasmuch as such a notion is independently plausible—or, at least, independently intelligible—such theorists carry the burden of explaining why such a notion appears to be intelligible, despite its ultimate unintelligibility. Along these lines, the theorist might argue that we are making a common, reasonable, but ultimately fallacious generalization from 'normal

[20] A paracomplete proposal rejects LEM, and a paraconsistent proposal rejects 'Explosion' (i.e., $\alpha, \neg\alpha \Rightarrow \beta$, in both Rule and Conditional form). (See Appendix for the former type of approach.)

cases' to all cases, or some such mistake. (E.g. some connective, if *restricted* to a proper fragment of our language, behaves in the EE way.) Alternatively, such theorists might argue that, contrary to initial appearances, the allegedly intelligible notion only appears to be a clear notion but, in fact, is rather unclear; once clarified, the alleged EE device (or whatever) is clearly not such a device. (E.g. one might argue that the alleged notion is a conflation of various notions, each one of which is intelligible but not one of which behaves in the alleged, problematic way.) Whatever the response, theorists do owe something to Rv3 revengers: an explanation as to why the given (and otherwise problematic) notion is unintelligible.

1.5 Some Closing Remarks

I have hardly scratched the surface of *revenge* in the foregoing remarks. The phenomenon (or, perhaps more accurately, family of phenomena) has in many respects been the fuel behind formal theories of truth, at least in the contemporary period. Despite such a role, a clear understanding of revenge is a pressing and open matter. What, exactly, is revenge? How, if at all, is it a serious problem? Is the problem *logical*? Is the problem *philosophical*? And relative to what end, exactly, is the alleged problem a problem? Answers to some of the given questions, I hope, are clear enough in foregoing remarks, but answers—clear answers—to many of the questions remain to be found. Until then, full evaluation of current theories of truth remains out of reach. The hope, however, is that the papers in this volume move matters forward.

Chapter Summaries

What follows are brief synopses of the chapters, ordered alphabetically in terms of author(s). The synopses are intended to help the reader find chapters of particular interest, rather than serve as discussion of the chapters.

Cook. Call a concept $\mathcal{C}$ *indefinitely extensible* just if there's a rule r such that, when applied to any 'definite collection' of objects falling under $\mathcal{C}$, r yields a new object falling under $\mathcal{C}$. In his 'Embracing revenge: on the indefinite extendibility of language', Roy Cook argues that the revenge phenomenon is reason to think that our concept of *language*, and the associated concept of *truth value* (or *semantic value*), is indefinitely extensible. In the end, the revenge phenomenon is a witness to the indefinite extensibility of our language, and, in particular, its 'semantic values'.

Eklund. In his 'The liar paradox, expressibility, possible languages', Matti Eklund focuses on general theses that are standardly tied to the Liar phenomenon. On one hand, there are two related lessons that are sometimes drawn from the Liar's revenge: namely, *radical inexpressibility* and (the weaker) *inexpressibility*. On the other hand, there are two related principles that often make for frustration in the face of revenge: namely, *semantic self-sufficiency* and (what Eklund calls) *weak universality*. Eklund elucidates the given theses, and focuses attention on inexpressibility and weak universality. Eklund argues that common approaches to such theses may confront difficulties from facts governing the space of possible languages, an issue at the heart of Eklund's essay.

Field. In his 'Solving the paradoxes, escaping revenge', Hartry Field advances a (paracomplete) theory of truth that, he argues, undermines the 'received wisdom' about revenge, where such wisdom, as Field puts it, maintains that 'any intuitively natural and consistent resolution of a class of semantic paradoxes immediately leads to other paradoxes just as bad as the first'. After presenting his own theory of truth (which extends the Kripke approach with a suitable conditional), Field argues that, pace 'received wisdom', it is revenge-free. The overall theory and arguments for its revenge-free status have provoked discussion in other chapters (see especially Leitgeb, Priest, Rayo–Welch).[21]

Hofweber. Validity is often thought to be truth-preserving: an inference rule is valid just if truth-preserving.[22] Thomas Hofweber, in his 'Validity, paradox, and the ideal of deductive logic', argues that two senses of 'an inference rule is valid just if truth-preserving' are important to distinguish. One sense is the 'strict reading', according to which *each and every* instance of the given rule is truth-preserving. The other reading is the 'generic reading', which, in some sense, is analogous to the claim that *bears are dangerous*, a claim that is true even though not true of all bears. This distinction, which Hofweber discusses, holds the key to resolving the revenge phenomenon. In

[21] One issue not discussed is the ideal of 'exhaustive characterization', according to which we can truly say (something equivalent to) that all sentences are either True, False, or Whathaveyou (where 'Whathaveyou' is a stand in for the predicates used to classify Liars or the like), and do as much in our own language. One might wonder whether the 'received wisdom' counts as 'natural' only those theories that afford exhaustive characterization, in which case, Field's argument against 'received wisdom' might miss the mark. (Without further clarification of 'exhaustive characterization', I do not intend these remarks as a serious objection, but rather only something for the reader to consider.)

[22] I should note that if, as is usual, 'truth-preserving' is understood via a conditional, so that $\langle \alpha, \beta \rangle$ is 'truth-preserving' just if $\alpha \rightarrow \beta$ is true (for some suitable conditional in the language), then many standard theories of *transparent truth* (i.e., fully intersubstitutable truth) will not have it that valid arguments are truth-preserving. See [2] for some discussion, but also [5] for broader, philosophical issues. This issue, regrettably, is not discussed much in this volume, but it is highly important. Restall's chapter (Chapter 12) has some direct relevance for the issue, as does Field's (Chapter 4). Hofweber briefly mentions the issue as it arises for Field's theory.

particular, the Liar's revenge teaches us that we should abandon the traditional ideal of deductive logic, which requires that our theories be underwritten by rules that are valid in the 'strict sense'. On the positive side, the Liar's revenge teaches us that we should embrace the 'generic' ideal of deductive logic, which requires only that our rules be 'generically valid'.

Leitgeb. In his 'On the metatheory of Field's *Solving the Paradoxes, Escaping Revenge*', Hannes Leitgeb argues that whether, in the end, Field's proposed theory escapes revenge turns on the details of its metatheory. Leitgeb argues that without a clear, explicitly formulated metatheory, the intended interpretation of Field's proposed truth theory—and, hence, the proposed resolution of paradox—remains unclear. What is ultimately required, Leitgeb argues, is a metatheory that includes a non-classical set theory for which the logic is the logic of Field's truth theory.[23] Towards moving matters forward, Leitgeb sketches two target metatheories, a classical and a non-classical one. Leitgeb conjectures that, for reasons he discusses, revenge may emerge for Field's proposal once a full metatheory is in place.

Maudlin. In his 'Reducing revenge to discomfort', Tim Maudlin argues that the revenge phenomenon ultimately teaches us something about our normative principles of assertion. As in §1.3 (above), invoking a new category for Liars—say, *bugger*—seems inevitably to lead to new Liars (e.g. the starred sentence in §1.3 above). While Maudlin maintains that we do need three semantic categories (viz., truth, falsity, and ungroundedness), he argues that we need no more than three. In particular, we may—and should—assert that the ticked sentence in §1.2 above is not true; it's just that we'll be bucking the traditional principle according to which only truths are properly assertible. The problem, Maudlin argues, is not with principles of truth (e.g. Release and Capture); the problem is with the traditional principle of assertion.[24] On the other hand, Maudlin admits that the revenge phenomenon returns even for his revised principle of assertion (e.g. 'I am not properly assertible according to Maudlin's revised principles'). Maudlin argues that this is revenge, but that it is at most a discomfort; it is far from threatening the coherence of *truth*.

Patterson. In his 'Understanding the liar', Douglas Patterson advances an 'inconsistency view' of the semantic paradoxes in English; however, his view is not a dialetheic

[23] Actually, Leitgeb's claim needn't be that the metatheory include a non-classical *set* theory, but rather that it include a non-classical theory of objects that play the relevant role that sets typically play—e.g. serving as a 'model' or etc.

[24] I should flag one potential confusion here. Maudlin claims, at least in his fuller work (see references in Chapter 7), that Rule Capture is *valid*, in the sense that, necessarily, if α is true, then so too is $Tr(\ulcorner\alpha\urcorner)$. At the same time, the logic governing assertibility is closer to *KF*, where $\alpha \Rightarrow Tr(\ulcorner\alpha\urcorner)$ fails in Rule form (and Conditional form).

view (according to which English is inconsistent, in the sense that some true English sentence has a true negation). Patterson argues that such a view is not that natural languages are inconsistent, but rather that competent speakers of natural languages process such languages in accord with an inconsistent theory. One of Patterson's principal aims is to show that, perhaps contrary to common thinking, *understanding* a language can be—and, in the case of English, is—a relation to a false theory. Patterson argues that such an 'inconsistency view' is the most promising lesson to draw from the revenge phenomenon.

Priest. In his 'Revenge, Field, and ZF', Graham Priest does three things. First, Priest characterizes the Liar's revenge, and carves up three options for dealing with it. Second, Priest directs the discussion towards Field's chapter (see Chapter 4), and argues that Field's proposal is not revenge-free, contrary to Field; in particular, it faces an expected problem with the notion of *having value 1*. (Priest anticipates the immediate thought that, as sketched in §1.4.1 above, he is merely launching a form of 'too-easy revenge', conflating model-dependent and 'real' notions. Priest argues that unless 'having value 1' is a real notion, Field has given no reason to think that Field's proposed logic has anything to do with real validity—i.e. validity in our real language.) Third, Priest argues that the (alleged) troubles facing Field's proposal are a symptom of deeper revenge in the background theory of ZF, which theory, Priest argues, itself faces a serious revenge-like situation involving V (the cumulative hierarchy): the logic defined by the theory (in terms of models) does not apply to the theory itself, thereby leaving us 'bereft of a justification for reasoning about sets', as Priest puts it.

Rayo and Welch. In their 'Field on revenge', Agustín Rayo and Philip Welch argue that Field's allegedly revenge-free truth theory (see Chapter 4) is not really revenge-free—or, at least, that its prospects for being revenge-free crucially depend on the outcome of current debates over higher-order languages. Rayo and Welch argue that, just as 'received wisdom' maintains, Field's proposed theory enjoys consistency only in virtue of expressive limitations. In particular, by invoking the appropriate higher-order language, we can explicitly characterize a key semantic notion involved in Field's proposal: viz., an *intended interpretation of* L^+, where L^+ is the language of Field's theory (a language enjoying transparent truth and a suitable conditional). Such a notion, as Rayo and Welch argue, plays the Liar's revenge role: it would generate inconsistency were it expressible in Field's proposed language.[25]

Read. In his 'Bradwardine's revenge', Stephen Read discusses a theory of truth proposed by Thomas Bradwardine (who was principally a physicist and theologian in

[25] I should be slightly more precise and note that Field (Chapter 4) considers a *class* of languages (or theories) that enjoy the desiderata of transparent truth and a suitable conditional, and Rayo and Welch direct their remarks against the relevant class.

the 1300s). Read shows that Bradwardine's theory, according to which Liars are not true (because they'd have to be true and not true, which is impossible), is a subtler theory than the later Buridan-like theories that, in effect, reject unrestricted Capture for truth (see §1.1 above). Moreover, the theory, on the surface, as Read argues, seems to promise a revenge-free approach to a whole host of semantic paradoxes. The key for Bradwardine is to distinguish between the claim that the Liar is false from the Liar itself. The propositions appear to be indistinguishable, but they are not. According to Bradwardine, any proposition that 'says' of itself that it is false, also 'says' of itself that it is true. (As Read points out, this is a subtler thesis than the later Buridanian claim that every claim 'says' of itself that it is true.)

Restall. In his 'Curry's revenge: the costs of non-classical solutions to the paradoxes of self-reference', Greg Restall discusses the challenges posed by Curry's paradox to those (non-classical) theories that attempt to preserve Capture and Release, in both Rule and Conditional forms, for truth and related semantic (or logical) notions—e.g. 'semantical properties', which serve as the 'extensions' of predicates in naïve semantics.[26] Restall argues that a Curry conditional is fairly easy to construct unless the language has fairly narrow limits. In particular, a theory that avoids Curry paradox must either reject 'large disjunctions', various (otherwise natural) forms of distribution, or the transitivity of entailment. As Restall notes, whatever option is rejected, sound philosophical motivation must accompany the rejection.

Scharp. In his wide-ranging 'Aletheic vengeance', Kevin Scharp argues that the Liar's revenge teaches us, among other things, that truth is an inconsistent concept the best theory of which implies that typical truth rules are 'constitutive' of truth but nonetheless invalid. Scharp argues that the best (inconsistency) theory of truth takes truth to be a *confused concept* (in a technical sense), but is a theory that does not *use* our concept of truth at all. Indeed, Scharp proposes that the proper approach to truth is one that finds other—non-confused—notions to play the truth role(s).

Shapiro. In his 'Burali-Forti's revenge', Stewart Shapiro turns the focus from the Liar paradox to the Burali-Forti paradox, which, he argues, has its own revenge issues. (Using the later von Neumann account, which came after Burali-Forti, the paradox, in short, is that the set Ω of all ordinals satisfies all that's required to be an ordinal, in which case, the successor of Ω, namely $\Omega + 1$, is strictly greater than Ω. But, being

[26] Restall doesn't use the term 'semantical properties', but he clearly has this under discussion. (Some philosophers refer to the target entities as 'naïve sets', but *sets* ultimately have little to do with the matter. If we let mathematicians tell us the 'nature' of *sets*—and they'll likely do so by axiomatizing away Russell problems—we still have to find a theory of 'semantical properties', the entities that play the familiar role in semantics, namely, those objects 'expressed' by any meaningful predicate and 'exemplified' by an object just if the given predicate is 'true of' the object.)

itself an ordinal, $\Omega + 1$ must be in Ω, giving the result that $\Omega < \Omega + 1 \leq \Omega$, which is impossible.) Shapiro presents the paradox and a variety of ways of dealing with it. He argues that each option faces severe problems, leaving the matter open.

Simmons. In his 'Revenge and context', Keith Simmons first distinguishes between (what he calls) *direct revenge* and *second-order revenge*. The former variety is the (what one might call 'first-order') variety: we already have a stock of semantic terms, and they generate paradox. In particular, as with the ticked sentence in §1.3, one is naturally inclined to classify it as 'neither true nor false', but this (at least prima facie) implies untruth, and the paradox remains. One is stuck in direct revenge: an inability to classify the sentence as one thinks it ought to be classified—but cannot be so classified, on pain of inconsistency. But, now, one introduces new, technical machinery to deal with the direct revenge problem: one calls the Liar a *bugger*, or *unstable*, or *whatnot*. Second-order revenge emerges with this new machinery, and one is, again, unable to classify the (new) Liars as they 'ought' to be (in some sense). Simmons argues that, while there is still work to be done, his 'singularity theory' of semantic notions deals not only with direct revenge in a natural way; it also holds the promise of resolving second-order revenge.

Appendix

Since many of the chapters in this volume presuppose familiarity with so-called fixed-point languages, and, in particular, *paracomplete* languages (see below), this appendix is intended as a user-friendly sketch of the (or a) basic background picture. In particular, I sketch a basic Kripkean picture [11], although I take liberties in the setting up.[27] I focus on the non-classical interpretation of Kripke's (least fixed point) account. My aim is only to give a basic philosophical picture and a *sketch* of the formal model. I focus on the semantic picture.

Philosophical picture

One conception of truth has it that truth is entirely *transparent*, that is, a truth predicate $Tr(x)$ in (and for) our language such that $Tr(\ulcorner\alpha\urcorner)$ and α are intersubstitutable in all (non-opaque) contexts, for all α in the language. This conception comes with a guiding metaphor, according to which 'true' is introduced only for purposes of generalization. Prior to introducing the device, we spoke only the 'true'-free fragment.

[27] This appendix is a very slightly altered version of a section from the much larger [2], which provides more references.

(Similarly for other semantic notions/devices, e.g., 'denotes', 'satisfies', 'true of', etc.) For simplicity, let us assume that the given 'semantic-free' fragment (hence, 'true'-free fragment) is such that LEM holds.[28] Letting $\mathcal{L}_0$ be our 'semantic-free fragment', we suppose that $\alpha \vee \neg\alpha$ is true for all α in $\mathcal{L}_0$.[29] Indeed, we may suppose that classical semantics—and logic, generally—is entirely appropriate for the fragment $\mathcal{L}_0$.

But now we want our generalization-device. How do we want this to work? As above, we want $Tr(\ulcorner\alpha\urcorner)$ and α to be intersubstitutable for *all* α. The trouble, of course, is that once 'is true' is introduced into the language, various unintended—and, given the role of the device, paradoxical—sentences emerge (e.g. the ticked sentence in §1.2 above).[30]

The *paracomplete* idea, of which Kripke's is the best known, is (in effect) to allow some instances of $\alpha \vee \neg\alpha$ to 'fail'.[31] In particular, if α itself fails to 'ground out' in $\mathcal{L}_0$, fails to 'find a value' by being ultimately equivalent to a sentence in $\mathcal{L}_0$, then the α-instance of LEM should fail. (This is the so-called *least fixed point* picture.)

Kripke illustrated the idea in terms of a learning or teaching process. The guiding principle is that $Tr(\ulcorner\alpha\urcorner)$ is to be asserted exactly when α is to be asserted. Consider an $\mathcal{L}_0$-sentence that you're prepared to assert—say, '$1 + 1 = 2$' or 'Max is a cat' or whatever. Heeding the guiding principle, you may then assert that '$1 + 1 = 2$' and 'Max is a cat' are true. In turn, since you are now prepared to assert

(1) 'Max is a cat' is true

the guiding principle instructs that you may also assert

(2) ' 'Max is a cat' is true' is true.

And so on. More generally, your learning can be seen as a process of achieving further and further truth-attributions to sentences that 'ground out' in $\mathcal{L}_0$. (Similarly for falsity, which is just truth of negation.) Eventually, your competence reflects precisely

[28] This assumption sets aside the issue of vagueness (and related sorites puzzles). I am setting this aside only for simplicity. The issue of vagueness—or, as some say, 'indeterminacy', in general—is quite relevant to some paracomplete approaches to truth. See [4], [13], [21].

[29] This assumption is not essential to Kripke's account; however, it makes the basic picture much easier to see.

[30] With respect to formal languages, the inevitability of such sentences is enshrined in Gödel's so-called *diagonal lemma*. (Even though the result is itself quite significant, it is standardly called a *lemma* because of its role in establishing Gödel–Tarski indefinability theorems. For user-friendly discussion of the limitative results, and for primary sources, see [18]. For a general discussion of diagonalization, see [19].)

[31] NB: The sense in which instances of $\alpha \vee \neg\alpha$ 'fails' is modeled by such instances being undesignated (in the formal model). (See 'Formal model' below.) How, if at all, such 'failure' is expressed *in* the given language is relevant to 'revenge', but I will leave chapters of this volume to discuss that.

the defining intersubstitutivity—and transparency—of truth: that $Tr(\ulcorner\alpha\urcorner)$ and α are intersubstitutable for *all* α of the language.

But your competence also reflects something else: namely, the failure to assert either α or $\neg\alpha$, for some α in the language. To see the point, think of the above process of 'further and further truth-attributions' as a process of writing two (very, very big) books—one, *The Truth*, the other *The False*. Think of each stage in the process as completing a 'chapter', with chapter zero of each book being empty—this indicating that *at the beginning* nothing is explicitly recorded as true (or, derivately, false).

Concentrate just on the process of recording *atomics* in *The Truth*. When you were first learning, you scanned $\mathcal{L}_0$ (semantic-free fragment) for the true (atomic) sentences, the sentences you were prepared to assert. Chapter one of *The Truth* comprises the results of your search—sentences such as 'Max is a cat' and the like. In other words, letting '$I(t)$' abbreviate *the denotation of t*, chapter one of *The Truth* contains all of those atomics $\alpha(t)$ such that $I(t)$ exemplifies α, a 'fact' that *would've been* recorded in chapter zero *had* chapter zero recorded the true semantic-free sentences. (For simplicity, if $\alpha(t)$ is an $\mathcal{L}_0$-atomic such that $I(t)$ exemplifies α, then we'll say that $I(t)$ exemplifies α *according to chapter zero*. In the case of 'Max is a cat', chapter zero has it that Max exemplifies cathood, even though neither 'Max is a cat' nor anything else appears in chapter zero.)

In the other book, *The False*, chapter zero is similarly empty; however, like chapter zero of *The Truth*, the sentences that *would* go into *The False*'s chapter zero are those (atomic) $\mathcal{L}_0$-sentences that, according to the world (as it were), are false—e.g., '$1 + 1 = 3$', 'Max is a dog', or the like.[32] If $\alpha(t)$ is a false $\mathcal{L}_0$-atomic, we'll say that *according to chapter zero*, $I(t)$ exemplifies $\neg\alpha$ (even though, as above, chapter zero explicitly records nothing at all). In turn, chapter one of *The False* contains all of those atomics $\alpha(t)$ such that, according to chapter zero, $I(t)$ exemplifies $\neg\alpha$ (i.e., the $\mathcal{L}_0$-atomics that are false, even though you wouldn't say as much at this stage).

And now the writing (of atomics) continues: chapter two of *The Truth* comprises 'first-degree truth-attributions' and atomics $\alpha(t)$ such that, as above, $I(t)$ exemplifies α according to chapter *one*, sentences like (1) and 'Max is a cat'. In turn, chapter three of *The Truth* comprises 'second-degree' attributions, such as (2), and atomics $\alpha(t)$ such that (as it were) t is α according to chapter *two*. And so on, and similarly for *The False*. In general, your writing-project exhibits a pattern. Where $I_i(Tr)$ is chapter i of *The Truth*, the pattern runs thus:

$$I_{i+1}(Tr) = I_i(Tr) \cup \{\alpha(t) : \alpha(t) \text{ is an atomic and } I(t) \text{ exemplifies } \alpha \text{ according to } I_i(Tr)\}$$

[32] For convenience, we'll also put non-sentences into *The False*. Putting non-sentences into *The False* is not essential to Kripke's construction, but it makes things easier. Obviously, one can't *write* a cat but, for present purposes, one can think of *The False* as a special book that comes equipped with attached nets (wherein non-sentences go), a net for each chapter.

Let $\mathcal{S}$ comprise all sentences of the language. With respect to *The False* book, the pattern of your writing (with respect to atomics) looks thus:

$$I_{i+1}(F) = I_i(F) \cup \{\alpha(t) : \alpha(t) \text{ is an atomic and } I(t) \notin \mathcal{S} \text{ or } I(t) \text{ exemplifies } \neg\alpha \text{ according to } I_i(Tr)\}$$

So goes the basic process for *atomics*. But what about compound (molecular) sentences? The details are sketched below (see 'Formal model'), but for now the basic idea is as follows (here skipping the relativizing to chapters). With respect to negations, $\neg\alpha$ goes into *The True* just when α goes into *The False*. (Otherwise, neither α nor $\neg\alpha$ finds a place in either book.) With respect to *conjunctions*, $\alpha \wedge \beta$ goes into *The False* if either α or β goes into *The False*, and it goes into *The True* just if both α and β go into *The True*. (Otherwise, $\alpha \wedge \beta$ finds a place in neither book.) The case of *disjunctions* is dual, and the quantifiers may be treated as 'generalized conjunction' (universal) and 'generalized disjunction' (existential). This approach to compound sentences reflects the so-called *Strong Kleene* scheme, which is given below (see 'Formal model').

Does every sentence eventually find a place in one book or other? No. Consider an atomic sentence λ, like the ticked sentence in §1.2, equivalent to $\neg Tr(\ulcorner\lambda\urcorner)$. In order to get λ into *The True* book, there'd have to be some chapter in which it appears. λ doesn't appear in chapter zero, since nothing does. Moreover, λ doesn't exemplify anything 'according to chapter zero', since chapter zero concerns only the $\mathcal{L}_0$-sentences (and λ isn't one of those). What about chapter one? In order for λ to appear in chapter one, λ would have to be in chapter zero or be such that λ exemplifies $\neg Tr(x)$ according to chapter zero. But for reasons just given, λ satisfies neither disjunct, and so doesn't appear in chapter one. The same is evident for chapter two, chapter three, and so on. Moreover, the same reasoning indicates that λ doesn't appear in *The False* book.

In general, Liar-like sentences such as the ticked sentence in §1.2 will find a place in one of our books only if it finds a place in one of the chapters $I_i(Tr)$ or $I_i(F)$. But the ticked sentence will find a place in $I_i(Tr)$ or $I_i(F)$ only if it finds a place in $I_{i-1}(Tr)$ or $I_{i-1}(F)$. But, again, the ticked sentence will find a place in $I_{i-1}(Tr)$ or $I_{i-1}(F)$ only if it finds a place in $I_{i-2}(Tr)$ or $I_{i-2}(F)$. And so on. But, then, since $I_0(Tr)$ and $I_0(F)$ are both empty, and since—by our stipulation—something exemplifies a property according to $I_0(Tr)$ only if the property is a non-semantic one (the predicate is in $\mathcal{L}_0$), the ticked sentence (or the like) fails to find a place in either book. Such a sentence, according to Kripke, is not only *ungrounded*, since it finds a place in neither book, but also *paradoxical*—it *couldn't* find a place in either book.[33]

[33] The force of *couldn't* here is made precise by the full semantics, but for present purposes one can think of *couldn't* along the lines of *on pain of (negation-) inconsistency* or, for that matter, *on pain of being in both books* (something impossible, on the current framework).

So goes the basic philosophical picture. What was wanted was an account of how, despite the existence of Liars, we could have a fully transparent truth predicate in the language—and do so without triviality (or, in Kripke's case, inconsistency). The foregoing picture suggests an answer, at least if we eventually have a chapter $I_i(Tr)$ such that $Tr(\ulcorner \alpha \urcorner)$ is in $I_i(Tr)$ if and only if α is in $I_i(Tr)$, and similarly a chapter for *The False*. What Kripke (and, independently, Martin–Woodruff) showed is that, provided our 'writing process' follows the right sort of scheme (in effect, a logic weaker than classical), our books will contain such target chapters, and in that respect our language can enjoy a (non-trivial, indeed consistent) transparent truth predicate. Making the philosophical picture more precise is the job of formal, philosophical modeling, to which I now briefly (and somewhat informally) turn.

Formal model

For present purposes, I focus on what is known as Kripke's 'least fixed point' model (with empty ground model). I leave proofs to cited works (all of which are readily available), and try to say just enough to see how the formal picture goes.

Following standard practice, we can think of an *interpreted language* $\mathcal{L}$ as a triple $\langle \mathsf{L}, \mathcal{M}, \sigma \rangle$, where L is the syntax (the relevant syntactical information), $\mathcal{M}$ is an 'interpretation' or 'model' that provides interpretations to the non-logical constants (names, function-symbols, predicates), and σ is a 'semantic scheme' or 'valuation scheme' that, in effect, provides interpretations (semantic values) to compound sentences.[34]

Consider, for example, familiar classical languages, where the set $\mathcal{V}$ of 'semantic values' is $\{1, 0\}$. In classical languages, $\mathcal{M} = \langle \mathcal{D}, I \rangle$, with $\mathcal{D}$ our (non-empty) domain and I an 'interpretation-function' that assigns to each name an element of $\mathcal{D}$ (the denotation of the name), assigns to each n-ary function-symbol an element of $\mathcal{D}^n \longrightarrow \mathcal{D}$, that is, an n-ary function from $\mathcal{D}^n$ into $\mathcal{D}$, and assigns to each n-ary predicate an element of $\mathcal{D}^n \longrightarrow \mathcal{V}$, a function—sometimes thought of as the *intension* of the predicate—taking n-tuples of $\mathcal{D}$ and yielding a 'semantic value' (a 'truth value'). The *extension* of an n-ary predicate F (intuitively, the set of things of which F is true) contains all n-tuples $\langle a_1, \ldots, a_n \rangle$ of $\mathcal{D}$ such that $I(F)(\langle a_1, \ldots, a_n \rangle) = 1$. The classical valuation scheme τ (for Tarski) is the familiar one according to which a negation is true (in a given model) exactly when its negatum is false (in the given model), a disjunction is true (in a model) iff one of the disjuncts is true (in the model), and existential sentences are treated as generalized disjunctions.[35]

[34] For present purposes, a semantic scheme or valuation scheme σ is simply some general definition of *truth (falsity) in a model*. For more involved discussion of semantic schemes, see [8].

[35] I assume familiarity with the basic classical picture, including 'true in $\mathcal{L}$' and so on. To make things easier, I will sometimes assume that we've moved to models in which everything in the domain has a name, and otherwise I'll assume familiarity with standard accounts of 'satisfies $\alpha(x)$ in $\mathcal{L}$'.

Classical languages (with suitably resourceful L) cannot have their own *transparent* truth predicate. Paracomplete languages reject the 'exhaustive' feature implicit in classical languages: namely, that a sentence or its negation is true, for *all* sentences.

The standard way of formalizing paracomplete languages expands the interpretation of predicates. Recall that in your 'writing process' some sentences (e.g. Liars) found a place in neither book. We need to make room for such sentences, and we can expand our semantic values $\mathcal{V}$ to do so; we can let $\mathcal{V} = \{1, \frac{1}{2}, 0\}$, letting the middle value represent (for 'modeling' purposes) the status of sentences that found a place in neither book.

Generalizing (but, now, straining) the metaphor, we can think of all *n*-ary predicates as tied to two such 'big books', one recording the objects of which the predicate is true, the other the objects of which it is false. On this picture, the *extension* of a predicate *F* remains as per the classical (containing all *n*-tuples of which the predicate is true), but we now also acknowledge an *antiextension*, this comprising all *n*-tuples of which the predicate is false. This broader picture of predicates enjoys the classical picture as a special case: namely, where we stipulate that, for any predicate, the extension and antiextension are jointly exhaustive (the union of the two equals the domain) and, of course, exclusive (the intersection of the two is empty).

Concentrating on the so-called *Strong Kleene* account [36],[36] the formal story runs as follows. We expand $\mathcal{V}$, as above, to be $\{1, \frac{1}{2}, 0\}$, and so our language $\mathcal{L}_\kappa = \langle \mathsf{L}, \mathcal{M}, \kappa \rangle$ is now a so-called three-valued language (because it uses three semantic values).[37] Our *designated values*—intuitively, the values in terms of which *validity* or *consequence* is defined—are a subset of our semantic values; in the Strong Kleene case, there is exactly one designated element, namely 1.

A (Strong Kleene) model $\mathcal{M} = \langle \mathcal{D}, I \rangle$ is much as before, with I doing exactly what it did in the classical case except that I now assigns to *n*-ary predicates elements of $D^n \longrightarrow \{1, \frac{1}{2}, 0\}$, since $\mathcal{V} = \{1, \frac{1}{2}, 0\}$. Accordingly, the 'intensions' of our paracomplete (Strong Kleene) predicates have three options: 1, $\frac{1}{2}$, and 0. What about *extensions*? As above, we want to treat predicates not just in terms of extensions (as in the

[36] This is one of the paracomplete languages for which Kripke proved his definability result. Martin–Woodruff proved a special case of Kripke's general 'fixed point' result, namely, the case for so-called 'maximal fixed points' of the *Weak Kleene* scheme, or *Weak Kleene languages*.

[37] Kripke [11] made much of emphasizing that 'the third value' is not to be understood as *a third truth value* or anything else other than 'undefined' (along the lines of Kleene's original work [10]). I will not make much of this here, although what to make of semantic values that appear in one's formal account is an important, philosophical issue. (Note that if one wants to avoid a three-valued language, one can let $\mathcal{V} = \{1, 0\}$ and proceed to construct a Kleene-language by using *partial functions* (hence, the standard terminology 'partial predicates') for interpretations. I think that this is ultimately merely terminological, but I won't dwell on the matter here.

classical languages) but also antiextensions. The *extension* of an n-ary predicate F, just as before, comprises all n-tuples $\langle a_1, \ldots, a_n \rangle$ of $\mathcal{D}$ such that $I(F)(\langle a_1, \ldots, a_n \rangle) = 1$. (Again, intuitively, this remains the set of objects of which F is true.) The *antiextension*, in turn, comprises all n-tuples $\langle a_1, \ldots, a_n \rangle$ of $\mathcal{D}$ such that $I(F)(\langle a_1, \ldots, a_n \rangle) = 0$. (Again, intuitively, this is the set of objects of which F is false.) Of course, as intended, an interpretation might fail to put x in either the extension or antiextension of F. In that case, we say (in our 'metalanguage') that, relative to the model, F is *undefined* for x.[38]

Letting $\mathcal{F}^+$ and $\mathcal{F}^-$ be the extension and antiextension of F, respectively, it is easy to see that, as noted above, classical languages are a special case of (Strong Kleene) paracomplete languages. Paracomplete languages typically eschew inconsistency, and so typically demand that $\mathcal{F}^+ \cap \mathcal{F}^- = \emptyset$, in other words, that nothing is in both the extension and antiextension of any predicate. In this way, paracomplete languages typically agree with classical languages. The difference, of course, is that paracomplete languages do *not* demand that $\mathcal{F}^+ \cup \mathcal{F}^- = \mathcal{D}$ for all predicates F. But paracomplete languages *allow* for such 'exhaustive constraints', and in that respect can enjoy classical languages as a special case.

To see the close relation between classical languages and Strong Kleene, notice that κ, the Strong Kleene valuation-scheme, runs as follows (here treating only $\neg$, $\vee$, and $\exists$). Where $V_{\mathcal{M}}(\alpha)$ is the semantic value of α in $\mathcal{M}$ (and, for simplicity, letting each object in the domain name itself), and, for purposes of specifying scheme κ, treating $\mathcal{V}$ as standardly (linearly) ordered:

K1. $V_{\mathcal{M}}(\neg\alpha) = 1 - V_{\mathcal{M}}(\alpha)$.
K2. $V_{\mathcal{M}}(\alpha \vee \beta) = \mathsf{max}(V_{\mathcal{M}}(\alpha), V_{\mathcal{M}}(\beta))$.
K3. $V_{\mathcal{M}}(\exists x\, \alpha(x)) = \mathsf{max}\{V_{\mathcal{M}}(\alpha(t/x)) : \text{for all } t \in \mathcal{D}\}$.

The extent to which classical logic is an extension of a given paracomplete logic depends on the semantic scheme of the language.[39] Since κ, as above, is entirely in keeping with the classical scheme *except* for 'adding an extra possibility', it is clear that every classical interpretation is a Strong Kleene-interpretation (but not vice versa).[40]

[38] A common way of speaking is to say that, for example, $F(t)$ is 'gappy' with respect to $I(t)$. This terminology is appropriate if one is clear on the relation between one's formal model and the target notions that the model is intended to serve (in one respect or other), but the terminology can also be confusing, since, e.g., in the current Strong Kleene language, one cannot truly assert of any α that α is 'gappy', i.e. $\neg Tr(\ulcorner \alpha \urcorner) \wedge \neg Tr(\ulcorner \neg\alpha \urcorner)$. (This issue arises in various chapters in the current volume.)

[39] Here, perhaps not altogether appropriately, I am privileging model theory over proof theory, thinking of 'logic' as the semantic consequence relation that falls out of the semantics. This is in keeping with the elementary aims of the essay, even though (admittedly) it blurs over a lot of philosophical and logical issues.

[40] Note that in classical languages, $V_{\mathcal{M}}(A) \in \{1, 0\}$ for any A, and the familiar classical clauses on connectives are simply (K1)–(K3).

Let us say that an interpretation *verifies* a sentence α iff α is designated (in this case, assigned 1) on that interpretation, and that an interpretation verifies a set of sentences Σ iff it verifies every element of Σ. We define *semantic consequence* in familiar terms: α is a consequence of Σ iff every interpretation that verifies Σ also verifies α. I will use '$\vdash_{SK}$' for the Strong Kleene consequence relation, so understood.

Let us say that a sentence α is logically true in $\mathcal{L}_\kappa$ exactly if $\emptyset \vdash_{SK} \alpha$, that is, iff α is designated (assigned 1) in every model. A remarkable feature of $\mathcal{L}_\kappa$ is that there are no logical truths. To see this, just consider an interpretation that assigns $\frac{1}{2}$ to every atomic, in which case, as an induction will show, every sentence is assigned $\frac{1}{2}$ on that interpretation. Hence, there's some interpretation in which no sentence is designated, and hence no sentence designated on all interpretations. A fortiori, LEM fails in Strong Kleene languages.[41]

And now an answer to one guiding question becomes apparent. What we want is a model of how our language can be non-trivial (indeed, consistent) while containing both a transparent truth predicate and Liar-like sentences. In large part, the answer is that our language is (in relevant respects) along Strong Kleene lines, that the logic is weaker than classical logic. Such a language, as Kripke showed, can contain its own (transparent) truth predicate.

The construction runs (in effect) along the lines of the 'big books' picture. For simplicity, let $\mathcal{L}_\kappa$ be a classical (but nonetheless Strong Kleene) language such that L (the basic syntax, etc.) is free of semantic terms but has the resources to describe its given syntax—including, among other things, having a name $\ulcorner\alpha\urcorner$ for each sentence α. (In other words, I assigns to each n-ary predicate an element of $\mathcal{D}^n \longrightarrow \{1, 0\}$, even though the values $\mathcal{V}$ of $\mathcal{L}_\kappa$ also contain $\frac{1}{2}$.) What we want to do is move to a richer language the syntax L^t of which contains $Tr(x)$, a unary predicate intended to be a transparent truth predicate for the enriched language. For simplicity, assume that the domain $\mathcal{D}$ of $\mathcal{L}_\kappa$ contains all sentences of L^t.[42]

Think, briefly, about the 'big books' picture. One can think of each successive 'chapter' as a language that expands one's official record of what is true (false). More formally, one can think of each such 'chapter' of both books as the extension and antiextension of 'true', with each such chapter expanding the interpretation of 'true'. Intuitively (with slight qualifications about chapters zero), one can think of $I_{i+1}(Tr)$ as explicitly recording *what is true according to chapter* $I_i(Tr)$. The goal, of course, is to find a 'chapter' at which we have $I_{i+1}(Tr) = I_i(Tr)$, a 'fixed point' at which anything

[41] This is not to say, of course, that one can't have a Strong Kleene—or, in general, paracomplete—language some proper fragment of which is such that $\alpha \vee \neg\alpha$ holds for all α *in the proper fragment*. (One might, e.g., stipulate that arithmetic is such that $\alpha \vee \neg\alpha$ holds.)

[42] This is usually put (more precisely) as that the domain contains the Gödel-codes of all such sentences, but for present purposes I will skip over the mathematical details.

true in the language is fully recorded in the given chapter—one needn't go further. Thinking of the various 'chapters' as *languages*, each with a richer interpretation of 'true', one can think of the 'fixed chapter' as a language that, finally, has a transparent truth predicate for itself.

Returning to the construction at hand, we have our Strong Kleene (but classical) 'ground language' $\mathcal{L}_\kappa$ that we now expand to $\mathcal{L}^t_\kappa$, the syntax of which includes that of $\mathcal{L}_\kappa$ but also has $Tr(x)$ (and the resulting sentences formable therefrom). We want the new language to 'expand' the ground language, and we want the former to have a model that differs from the latter only in that it assigns an interpretation to $Tr(x)$. For present purposes, we let I^t, the interpretation function in $\mathcal{L}^t_\kappa$, assign $(\emptyset, \emptyset)$ to $Tr(x)$, where $(\emptyset, \emptyset)$ is the function that assigns $\frac{1}{2}$ to each element of $\mathcal{D}^t$. (Hence, the extension and antiextension of $Tr(x)$ in $\mathcal{L}^t_\kappa$ are both empty.) This is the formal analogue of 'chapter zero'.

The crucial question, of course, concerns *further expansion*. How do we expand the interpretation of $Tr(x)$? How do we move to 'other chapters'? How, in short, do we eventually reach a 'chapter' or language in which we have a transparent truth predicate for the whole given language? This is the role of Kripke's 'jump operator'. What we want, of course, are 'increasingly informative' interpretations $(\mathcal{T}_i^+, \mathcal{T}_i^-)$ of $Tr(x)$, but interpretations that not only 'expand' the previous interpretations but also *preserve* what has already been interpreted. If α is true according to chapter i, then we want as much preserved: that α remain true according to chapter $i+1$. This is the role of the 'jump operator', a role that is achievable given the so-called *monotonicity* of Strong Kleene valuation scheme κ.[43] The role of the jump operator is to eventually 'jump' through successive interpretations (chapters, languages) $I_i(Tr)$ and land on one that serves the role of transparent truth—serves as an interpretation of 'is true'. As above, letting $I_i(Tr)$ be a function $(\mathcal{T}_i^+, \mathcal{T}_i^-)$ yielding 'both chapters i', the goal is to eventually 'jump' upon an interpretation $(\mathcal{T}_i^+, \mathcal{T}_i^-)$ such that $(\mathcal{T}_i^+, \mathcal{T}_i^-) = (\mathcal{T}_{i+1}^+, \mathcal{T}_{i+1}^-)$.

Focusing on the 'least such point' in the Strong Kleene setting, Kripke's construction proceeds as above. We begin at stage 0 at which $Tr(x)$ is interpreted as $(\emptyset, \emptyset)$, and we

[43] Monotonicity is the crucial ingredient in Kripke's (similarly, Martin–Woodruff's) general result. Let $\mathcal{M}$ and $\mathcal{M}'$ be paracomplete (partial) models for (uninterpreted) L. Let $\mathcal{F}_M^+$ be the extension of F in $\mathcal{M}$, and similarly $\mathcal{F}_{M'}^+$ for $\mathcal{M}'$. (Similarly for antiextension.) Then $\mathcal{M}'$ *extends* $\mathcal{M}$ iff the models have the same domain, agree on interpretations of names and function signs, and $\mathcal{F}_M^+ \subseteq \mathcal{F}_{M'}^+$ and $\mathcal{F}_M^- \subseteq \mathcal{F}_{M'}^-$ for all predicates F that M and M' interpret. (In other words, $\mathcal{M}'$ doesn't change $\mathcal{M}$'s interpretation; it simply interprets whatever, if anything, $\mathcal{M}$ left uninterpreted.) Monotonicity property: A semantic (valuation) scheme σ is *monotone* iff for any α that is interpreted by both models, α's being designated in $\mathcal{M}$ implies its being designated in $\mathcal{M}'$ whenever $\mathcal{M}'$ extends $\mathcal{M}$. So, the montonicity property of a scheme ensures that it 'preserves truth (falsity)' of 'prior interpretations' in the desired fashion.

define a 'jump operator' on such interpretations:[44] *Tr*(*x*) is interpreted as $(\mathcal{T}^+_{i+1}, \mathcal{T}^-_{i+1})$ at stage $i+1$ if interpreted as $(\mathcal{T}^+_i, \mathcal{T}^-_i)$ at the preceding stage i, where, note well, $\mathcal{T}^+_{i+1}$ comprises the sentences that are true (designated) at the preceding stage (chapter, language) i, and $\mathcal{T}^-_{i+1}$ the false sentences (and, for simplicity, non-sentences) at i. Accordingly, we define the 'jump operator' J_{SK} thus:[45]

$$J_{SK}(\mathcal{T}^+_i, \mathcal{T}^-_i) = (\mathcal{T}^+_{i+1}, \mathcal{T}^-_{i+1})$$

The jump operator yields a sequence of richer and richer interpretations that 'preserve prior information' (given monotonicity), a process that can be extended into the transfinite to yield a sequence

$$(\mathcal{T}^+_0, \mathcal{T}^-_0), (\mathcal{T}^+_1, \mathcal{T}^-_1), \ldots, (\mathcal{T}^+_\gamma, \mathcal{T}^-_\gamma), \ldots$$

defined (via transfinite recursion) thus:[46]

Jb. Base. $(\mathcal{T}^+_0, \mathcal{T}^-_0) = (\emptyset, \emptyset)$.

Js. Successor. $(\mathcal{T}^+_{\gamma+1}, \mathcal{T}^-_{\gamma+1}) = J_{SK}((\mathcal{T}^+_\gamma, \mathcal{T}^-_\gamma))$.

Jl. Limit. For limit stages, we collect up by unionizing the prior stages:

$$(\mathcal{T}^+_\lambda, \mathcal{T}^-_\lambda) = \left(\bigcup_{\varepsilon<\lambda} \mathcal{T}^+_\varepsilon, \bigcup_{\varepsilon<\lambda} \mathcal{T}^-_\varepsilon\right)$$

What Kripke showed—for *any* monotone scheme, and a fortiori for Strong Kleene—is that the transfinite sequence reaches a stage at which the desired transparent truth predicate is found, a 'fixed point' of the jump operator such that we obtain

$$(\mathcal{T}^+_\gamma, \mathcal{T}^-_\gamma) = (\mathcal{T}^+_{\gamma+1}, \mathcal{T}^-_{\gamma+1}) = J_{SK}((\mathcal{T}^+_\gamma, \mathcal{T}^-_\gamma))$$

The upshot is that 'chapter γ' or 'language γ' is such that $\mathcal{T}^+_\gamma$ and $\mathcal{T}^-_\gamma$ comprise all of the true (respectively, false) sentences of $\mathcal{L}^\gamma_k$, the Strong Kleene language at γ, which is to say that $\mathcal{L}^\gamma_k$ contains its own transparent truth predicate.

[44] So, our operator operates on the set of all (admissible) functions from $\mathcal{D}$ into $\{1, \frac{1}{2}, 0\}$, where $\mathcal{D}$ is in our given 'ground language'.

[45] Note that Kripke's definition applies to *any* monotone scheme σ. I relativize the operator to *SK* just to remind that we here focusing on the Strong Kleene case.

[46] Transfinite recursion is much like ordinary recursive definitions except for requiring an extra clause for so-called 'limit ordinals'. Here, γ and ε are ordinals (not sentences!), and λ a 'limit ordinal' (not a Liar!). (One can find a discussion of transfinite recursion in most standard set theory books or metatheory textbooks. Additionally, [13] and [19] are very useful, with the former especially useful for the present applications.)

The *proof* of Kripke's result is left to other (widely available) work.[47] Comments on the *adequacy* of Kripke's proposal may be found in some of the chapters in this volume, and in many of the cited works in any of the chapters.[48]

References

[1] Beall, JC (2006a). 'True, false, and paranormal'. *Analysis* 66(2): 102–14. Available in *Analysis Preprint* series

[2] ——— (2006b). 'Truth and paradox: a philosophical sketch'. In Dale Jacquette, (ed.), *Philosophy of Logic, Handbook of Philosophy of Science*, under the general editorship of Dov Gabbay, Paul Thagard, and John Woods), pp. 325–410. Elsevier, Dordrecht

[3] Feferman, Solomon (1984). 'Toward useful type-free theories, I'. *Journal of Symbolic Logic* 49: 75–111. Reprinted in Martin (1984)

[4] Field, Hartry (2003). 'The semantic paradoxes and the paradoxes of vagueness'. In JC Beall, (ed.), *Liars and Heaps: New Essays on Paradox*, pp. 262–311. Oxford University Press, Oxford

[5] ——— (2006). 'Truth and the unprovability of consistency'. *Mind*, 115: 567–606

[6] Fitting, Melvin (1986). 'Notes on the mathematical aspects of Kripke's theory of truth'. *Notre Dame Journal of Formal Logic* 27: 75–86

[7] Gupta, Anil (1997). 'Definition and revision'. In Enrique Villanueva, (ed.) *Truth*, number 8 in Philosophical Issues, pp. 419–43. Ridgeview Publishing Company, Atascadero, Calif.

[8] Gupta, Anil, and Belnap, Nuel (1993). *The Revision Theory or Truth*, MIT Press

[9] Jacquette, Dale (2004). 'Diagonalization in logic and mathematics'. In Dov M. Gabbay and Franz Günthner (eds.), *Handbook of Philosophical Logic*, pp. 55–147. Kluwer Academic Publishers, Dordrecht, 2nd edn.

[10] Kleene, S. C. (1952). *Introduction to Metamethematics*. North-Holland

[11] Kripke, Saul (1975). 'Outline of a theory of truth'. *Journal of Philosophy* 72: 690–716. Reprinted in Martin (1984)

[47] Kripke's own proof is elegant, bringing in mathematically important and interesting results of recursion theory and inductive definitions. Kripke's proof is also perhaps more philosophically informative than a popular algebraic proof, especially with respect to the least fixed point (on which we've focused here). Still, if one simply wants *a* proof of the given result (e.g., existence of least fixed point), a straightforward algebraic proof is available, due to Visser [22] and Fitting [6], and discussed in a general, user-friendly fashion by Gupta–Belnap [8].

[48] Acknowledgements: I am grateful to Colin Caret, Hartry Field, Lionel Shapiro, and Josh Schechter for comments and discussion.

[12] Martin, Robert L. (1984) (ed.), *Recent Essays on Truth and the Liar Paradox*. Oxford University Press, New York

[13] McGee, Vann (1991). *Truth, Vagueness, and Paradox*. Hackett, Indianapolis

[14] Priest, Graham (2006). *Doubt Truth to be a Liar*. Oxford University Press, Oxford

[15] ——— (2006). *In Contradiction*. Oxford University Press, Oxford, 2nd edn.

[16] Reinhardt, William N. (1986). 'Some remarks on extending and interpreting theories with a partial predicate for truth'. *Journal of Philosophical Logic* 15: 219–51

[17] Skyrms, Brian (1970). 'Return to the liar: three-valued logic and the concept of truth'. *American Philosophical Quarterly* 7: 153–61

[18] Smullyan, Raymond M. (1992). *Gödel's Incompleteness Theorems*, vol. 19 of *Oxford Logic Guides*. Oxford University Press, New York

[19] ——— (1993). *Recursion Theory for Metamathematics*, vol. 20 of *Oxford Logic Guides*. Oxford University Press, New York

[20] ——— (2001). 'Gödel's incompleteness theorems'. in Lou Goble (ed.), *The Blackwell Guide to Philosophical Logic*, pp. 72–89. Blackwell, Oxford

[21] Soames, Scott (1999). *Understanding Truth*. Oxford University Press, New York

[22] Visser, Albert (2004). 'Semantics and the liar paradox'. In Dov M. Gabbay and Franz Günthner, (eds.), *Handbook of Philosophical Logic*, pp. 149–240. Kluwer Academic Publishers, Dordrecht, 2nd edn.

[23] Yablo, Stephen (2003). 'New grounds for naïve truth theory', in JC Beall (ed.), *Liars and Heaps: New Essays on Paradox*, pp. 313–30. Oxford University Press, Oxford

2

Embracing Revenge: On the Indefinite Extendibility of Language

Roy T. Cook

2.1 Liars and Stronger Liars

Assume, as seems to be the case,[1] that our language contains the resources to meaningfully self-refer. In other words, locutions such as:

This sentence is false.

or:

Λ: Λ is false.

are meaningful.[2] If so, then we seemed forced to give up one of the following apparent platitudes:

Biv: Every sentence is either true or false
LNC: No sentence is both true and false
Truth: A sentence Φ is true if and only if what it says is the case.

[1] And as Saul Kripke has forcefully argued. See Kripke (1975), p. 56.

[2] Actually, this way of formulating things obscures some important complications. For example, if it is propositions, and not sentences, that are the primary bearers of truth, falsity, and the rest, then the sentence 'This sentence is false' would constitute, not a paradox, but a category mistake (thanks to an anonymous referee for pointing out the importance of this). Nevertheless, for the sake of simplicity I will continue to talk of sentences having truth values. Nothing important hinges on this, however, and a version of the view being defended here (as well as the paradoxes that motivate the view) can be reformulated quite simply in terms of propositions if the reader finds such talk more comfortable.

The argument runs as follows:

By *Biv*, Λ must be either true or false.

Assume Λ is true. Then, by *Truth*, what Λ says must be the case. Λ says that Λ is false. So Λ must be false. Thus, Λ is both true and false, contradicting *LNC*. So, contrary to our assumption, Λ cannot be true.

Assume that Λ is false. Λ says that Λ is false. So what Λ says is the case. So, by *Truth*, Λ is true. Thus, Λ is both true and false, again contradicting *LNC*. So, contrary to our assumption, Λ cannot be false.

Thus, the existence and meaningfulness of the Liar sentence Λ seems to force us to abandon (at least) one of *Biv, LNC*, or *Truth*.

Here we shall not consider giving up *Truth. Truth* is not a substantial theory regarding the nature of truth or our understanding of it, but instead merely reports the simple fact that the truth of a sentence depends upon the relationship between how the sentence represents the world as being and how the world really is. *Truth* makes no claim to tell us anything regarding what that relationship amounts to, nor does it take a stand on how substantial this relationship is. While those fond of *Biv* and *LNC* might be able to formulate an alternative notion of truth that did not respect *Truth*, we will not entertain such ideas here.[3]

Thus, our only option is to jettison either *Biv* or *LNC*. Either option, in the end, amounts to admitting a new category into our semantic classification of sentences. Before the paradox, sentences in our language appeared to be separable into two exclusive and exhaustive classes—the true and the false. Now, some sentences (such as the *Liar*) fall into neither category. In order to be as neutral as possible we shall call such sentences *pathological*.[4]

So, each sentence falls into exactly one of three categories: the true, the false, and the pathological. We can rephrase *Biv* and *LNC* as follows (*Truth* remains unchanged):

> *Biv**: Every sentence is either true, false, or pathological.
> *LNC**: No sentence is more than one of true, false, or pathological.

Everyday sentences (e.g., those not involving certain sorts of fixed point) will still fall into one or the other of our original categories, and problematic sentences such as the Liar are classified as pathological.

[3] The interested reader should consult Stephen Read's contribution to the present volume, which expresses such doubts regarding *Truth*.

[4] Note that the label 'pathological' is merely a placeholder for whatever substantial account we eventually give such problematic sentences (see the next section), and should not be interpreted as suggesting that such sentences are meaningless, that they fail to assert what they appear to assert, or anything else.

At this point, however, the Liar's younger and stronger cousin makes its appearance. If we allow a third semantic category into our account, then presumably there is nothing to stop us from adding to our language the expressive resources needed to talk about that category. In other words, just as we can talk about sentences being true or false, we can talk about sentences being pathological. But then we can formulate the *Strengthened Liar*:

Σ: Σ is either false or pathological.

The argument to contradiction is merely a slightly more complicated version of the argument for the Liar:

By *Biv*, Σ must be either true, false, or pathological.

Assume Σ is true. Then, by *Truth*, what Σ says must be the case. Σ says that Σ is either false or pathological. So Σ must be either false or pathological. Thus, Σ is either both true and false, or both true and pathological. Either option contradicts *LNC**. So, contrary to our assumption, Σ cannot be true.

Assume that Σ is false. So Σ is either false or pathological. Σ says that Σ is either false or pathological. So what Σ says is the case. So, by *Truth*, Σ is true. Thus, Σ is true and false, contradicting *LNC**. So, contrary to our assumption, Σ cannot be false.

Assume that Σ is pathological. So Σ is either false or pathological. Σ says that Σ is either false or pathological. So what Σ says is the case. So, by *Truth*, Σ is true. Thus, Σ is true and pathological, contradicting *LNC**. So, contrary to our assumption, Σ cannot be pathological.

Along the same lines as before, we seem forced to give up either *Biv** or *LNC** (assuming, again, that abandoning *Truth* is not an option). To do so amounts to adding a fourth semantic category, pathological$_2$, under which the Strengthened Liar and its ilk would fall.

But then we could formulate the *Strengthened-Strengthened Liar*:

$\Sigma\Sigma$: $\Sigma\Sigma$ is either false or pathological or pathological$_2$.

An argument similar to that above leads us to adding a fifth semantic value, and so on *ad infinitum*.

This particular infinite regress is just one instance of a more general problem plaguing solutions to the semantic paradoxes: *The Revenge Problem*. The Revenge Problem, simply put, is just this: Given any account that purports to deal adequately with a particular paradox, that account will rely on concepts (such as 'is pathological' above) which, if allowed into the object language, generate new paradoxes that cannot be dissolved by the account in question.

Revenge Problems affect just about any proposed solution to the semantic paradoxes, and the most prevalent response is just to deny that the concepts in question are, in fact, expressible in the object language, either by invoking a hierarchy of metalanguages, as in Tarski (1933), or by denying that the notions are expressible at

all. This approach violates strong intuitions regarding the functioning of language and truth, however. In particular, it appears as if we can meaningfully (if mistakenly) say things such as 'All sentences of this language are true'. As a result, an account of semantic paradox which minimized restrictions on what can be said would be preferable.

Given such generosity with regard to what we can express, we seem stuck with Revenge. The question, then, is what to do about it? The answer, which I intend to defend in the remainder of the chapter, is just this: Embrace Revenge. The Revenge Problem, it turns out, is not a problem at all, but instead affords the crucial insight that allows for a truly satisfactory solution to the semantic paradoxes.

Let us look a bit more closely at the phenomenon sketched above. We start with a language L_0 (say, first-order arithmetic or set theory) for which a classical, two-valued semantics suffices. If we extend the language to a new language L_1 by adding the expressive resources necessary to describe that semantic theory, however (i.e., we add a truth predicate), then we find that two truth values no longer suffice (thanks to the Liar). Once we have extended the language, we seemed forced to extend the semantics. The most straightforward way to do so it to add a third truth-value (our pathological above). With the new value we can consistently interpret L_1, and furthermore, we can describe the semantics for L_0 within L_1. If we attempt to describe the semantics of L_1, we discover a problem—the expressive resources of L_1 are not up to the task.[5] So we extend the language again (adding, e.g. an 'is pathological' predicate) obtaining a new language L_2. Our three-valued semantics does not allow a consistent interpretation of L_2 (thanks to the Strengthened Liar), so we are forced to extend the semantics by adding another truth-value (pathological_2), and so on. . .

At this point most of us stop, and find some other approach for dealing with the paradoxes. Instead, we should embrace the phenomenon: If three values do not suffice for L_2, then we should allow for a fourth truth-value to deal with the problematic sentences. If we then extend the language again, and four values no longer suffice, then add a fifth value. And then, if we extend the language again, a sixth truth-value emerges. In the end, there will be an infinitely ascending sequence of languages, and a corresponding infinitely ascending sequence of truth-values—one to handle each new Revenge Problem.

Of course, the advice to just keep adding truth-values as they are needed is worth little without some philosophical justification or story to back it up. The back story in the present case is this: The problems associated with the Liar paradox and similar

[5] Note, however, that the resources of L_1 do allow for a partial account of the semantics of L_1 within L_1 itself. In particular, L_1 can contain its own truth-predicate T(x) where Φ and $T(\langle \Phi \rangle)$ are everywhere intersubstitutable. The desire to retain this sort of uniform, transparent truth predicate motivates much of the present proposal.

phenomena can be traced to the fact that language is, in principle, indefinitely extensible. As a result of this particular sort of indefiniteness, we can never speak a 'universal langauge', quantify over all sentences, or speak of all truth-values at once. The semantic paradoxes, on this picture, are thereby merely attempts to express what cannot be (consistently) expressed (in any language).

The first description of what has (following Dummett) come to be called indefinite extensibility occurs in the writings of Russell on the paradox that bears his name:

> . . . the contradictions result from the fact that . . . there are what we may call *self-reproductive* processes and classes. That is, there are some properties such that, given any class of terms all having such a property, we can always define a new term also having the property in question. Hence we can never collect *all* of the terms having the said property into a whole; because, whenever we hope we have them all, the collection which we have immediately proceeds to generate a new term also having the said property.
>
> (1906, p. 144)

The idea that many mathematical concepts (such as set, ordinal, cardinal, etc.) are indefinitely extensible, and that many paradoxes are caused by a failure to recognize this (especially the set-theoretic paradoxes), is now a largely unchallenged part of the philosophical folklore.

For example, the concept *ordinal*[6] is indefinitely extensible since we can formulate a rule which, when applied to any definite collection of ordinals (i.e. any set of ordinals), 'generates' a new ordinal not contained in that collection (e.g. the transitive closure of the set of ordinals in question). As a result, there can be no definite collection of 'all' ordinals, since applying the rule to such a collection provides us with an ordinal that is not in this (supposedly exhaustive) collection. This is the real lesson of the Burali–Forti paradox. Similarly, since there can be neither a set of all sets, nor (on standard set-theoretic accounts of cardinal number) a cardinal number of the 'collection' of all cardinals, these notions are indefinitely extensible as well.

The idea that the semantic paradoxes are connected to indefinite extensibility of some sort, on the other hand, is far from accepted wisdom. The view is not completely new, however—Michael Dummett suggests just such a view in *The Seas Of Language*:

> The paradoxes—both the set theoretic and *the semantic paradoxes*—result from our possessing indefinitely extensible concepts. . . An indefinitely extensible concept is one for which, together with some determinate range of ranges of objects falling under it, we are given an intuitive principle whereby, if we have a sufficiently definite grasp of any one such range of objects, we

[6] Note that we say that the concept *ordinal* is indefinitely extensible, and not that the ordinals themselves are, since there is at least some controversy regarding whether we can quantify over the entirety of the extension of any indefinitely extensible concept.

can form, in terms of it, a conception of a more inclusive such range. . . By the nature of the case, we can form no clear conception of the extension of an indefinitely extensible concept; any attempt to do so is liable to lead us into contradiction.

(Dummett (1993), p. 454, emphasis added).

For Dummett, the semantic paradoxes can be traced to the fact that the concept *statement* is indefinitely extensible. Unfortunately, Dummett's solution—adopt intuitionistic logic when dealing with indefinitely extensible concepts—is clearly insufficient (Williamson (1998) points this out, in what appears to be the only sustained examination of Dummett's position on the semantic paradoxes).[7] Liar sentences are easy to construct within intuitionistic theories, and the constructivist constraints on inference adopted in such theories do nothing to block the derivation of a contradiction.[8]

The remainder of this chapter is devoted to exploring a variation of this idea of Dummett's. Instead of adopting intuitionistic logic in the face of indefiniteness, however, a different approach will be taken (one more in line with the traditional response to the indefinite extensibility of the set-theoretic hierarchy). The crucial claim, on which the rest depends, is the idea that it is not just the concepts *statement* or *language*, but also the concept *truth-value*, that is indefinitely extensible. Thus, we need first to examine exactly what a truth-value is.

2.2 What are Truth-Values?

Philosophers sometimes talk as if sentences are true solely because truth-values attach to them in the appropriate way (for example, explaining the (classical) logical truth of $P \vee \neg P$ in terms of the fact that every row of its truth table receives a T). Even if truth-values are legitimate objects, however (a view Frege held, but which has since become at least somewhat controversial), it is obvious that this is not the whole story. Rather, a truth-value attaches to a particular sentence because that sentence has the appropriate relationship to the world (where the phrase 'the world' can be interpreted as broadly as we need it to be, encompassing not just empirical matters, but semantic ones as well). Consider again the platitude *Truth*:

Truth: A sentence is true if and only if what it says is the case.

[7] Unlike the present approach, however, Williamson eventually abandons the idea that semantic paradoxes are linked to indefinite extensibility, opting instead for a view where the extension of the truth predicate is in a certain sense vague (see Williamson (1998), pp. 20–1).

[8] Of course, Dummett also (infamously) argues for the indefinite extensibility of the natural numbers. Since we are here citing Dummett merely for inspiration and general philosophical motivation, and will be departing from his philosophical picture in most of the details, we can ignore the less convenient aspects of his view.

Truth tells us that a sentence receives the value true just in case the state of affairs described by the sentence in question actually obtains in the world (however we flesh this out in the end). A similar platitude handles sentences that receive the value false:

Falsity : A sentence is false if and only if what it says fails to be the case.

Thus, truth-values are merely proxies for the sorts of relation that might hold between a sentence and the world.

Once we realize that truth-values are, in the final analysis, merely a means for keeping track of various relationships that hold between a sentence and the world, we can now consider how many truth-values there are, and how they attach to sentences in various languages. The crux of the matter, of course, is this: Given a language L, the appropriate semantics for L will contain exactly as many truth-values as there are ways for sentences of L to relate to the world.

For languages with no semantic vocabulary (such as first-order arithmetic or set theory), the two classical values suffice, since their simplicity guarantees that only two possible relationships between sentence and world are possible: either what a sentence says is the case, or what it says is not the case. Thus, once each atomic sentence is assigned a value of either true or false, every other sentence receives one of these two values in an orderly manner in accordance with *Truth* and *Falsity*.

Once we add semantic vocabulary to the language, however, we can no longer rely on two values to do the job. While all sentences from the original, unextended language will still receive one of the two original truth-values, the addition of semantic vocabulary such as a truth predicate allows for the construction of new sentences (such as the Liar) which cannot fall into either the 'what it says is the case' category or the 'what it says is not the case' category. Instead, to put it a bit loosely, since the Liar Sentence is true if and only if it is false, it falls into the 'what is says is the case if and only if what it says is not the case' category. So, in extending the expressive resources of our language, we find that there is now a new relationship that sentences can have to the world, and with it, a new truth-value.[9]

Upon reflection, however, this should not surprise us. If truth-values are the result of different sorts of relationships that sentences can have to the world, then the number of different sorts of such relationships will co-vary with two things: the complexity of the sentences, and the complexity of the world. Since adding new semantic vocabulary to our language potentially increases both the complexity of the sentences in our language and the complexity of the world (since linguistic facts are part of the world), it is perfectly reasonable that such extensions can introduce new

[9] Of course, this new truth-value is not 'new' in the sense that it did not exist prior to the extension of our language (whatever we mean by 'exist' here). Rather, it only applies to sentences of the new, extended language.

sorts of relations between the sentences of our language and the world. To sum all this up, more expressively powerful languages (in particular, those that contain more semantic vocabulary) require more truth-values.

In order to illustrate further how extending the expressive resources of a language can complicate the sort of relationship that a sentence can have to the world, and thus add additional truth-values, let us return to our earlier example. At this stage we have a language that contains a truth predicate, and three truth-values: true, false, and pathological. We can provide a platitude governing pathological sentences along something like the following lines:

> Path : A sentence is pathological if and only if what is says is the case if and only if what it says is not the case.

Suppose, however, that we extend the resources of our language again, by adding an 'is pathological' predicate. We now find that three truth-values are not enough, given that we can formulate sentences such as the Strengthened Liar. The Strengthened Liar cannot fall into either the 'what it says is the case' category or the 'what it says is not the case' category, but it also cannot fall into the 'if what is says is the case if and only if what it says is not the case' category. Rather, the Strengthened Liar is true if and only if it is either false or pathological, so the relation that holds between it and the world is something like 'what is says is the case if and only if either what it says is not the case or what it says is the case if and only if what it says is not the case'. Since 'what it says is the case' and 'what it says is not the case' are no longer exhaustive, this complicated relation is distinct from the relations between language and world corresponding to truth, falsity, or pathologicality. So we have a fourth truth-value.

We can continue, of course, by providing this new truth-value with a name, and then extending the language by adding the resources to refer to it. We then obtain a new *Strengthened-Strengthened Liar*, and a fifth truth-value. In the end, there will be an infinite series of languages, each more expressively robust than the one before it, but each also requiring a more complex semantics than the one before it.

The picture of language just sketched is the correct lesson to take from the Revenge Problem. The Revenge Problem shows us that, given any language L, L cannot contain the expressive resources necessary to completely describe the semantics appropriate for L. This is compatible with there being, for any language L, an extension of L, L', which is expressively rich enough to describe the semantics of L. What is required is an infinite sequence of languages, each one expressively richer than the last, and an infinite sequence of semantics, one for each language. The next section provides the formal details of such an account.

Note that all of the above can be summed up quite nicely as follows: The concept truth-value is indefinitely extensible. Recall that a concept is indefinitely extensible

if and only if there is a rule such that, given any definite collection of objects falling under the concept, application of the rule to that collection provides a new object also falling under the concept. The Revenge Problem, properly understood, provides just such a rule. Given a definite language L, let TV_L be the collection of truth-values required to correctly interpret L. Then the appropriate version of the Strengthened Liar (i.e. 'This sentence has a truth-value in TV_L other than true.') provides a truth-value which is not in TV_L (namely, the truth-value of that very sentence, the 'next' pathological value).

This completes the philosophical motivation of the proposed story (and solution to the Revenge Problem, which is now no longer a problem at all, but a feature of the theory). In the next section I provide a sketch of a formal development of the view, and after that I address some potential objections. The final section makes some quick observations regarding the conditional.

2.3 The Formal Theory

We begin with a sequence of languages L_α for α any ordinal.[10] L_0 contains (at least) the vocabulary of first-order ZFC with the usual formation rules (in what follows the finite ordinals are natural numbers, and their names are numerals), and its logical vocabulary consists of disjunction ($\vee$), conjunction ($\wedge$), negation ($\neg$), and the universal and existential quantifiers ($\forall$, $\exists$). L_1 contains L_0 plus a truth predicate 'T(x)' and a falsity predicate 'F(x)'.[11] The vocabulary of L_α, for $\alpha \geq 2$, is defined recursively as follows:

Successor:

The vocabulary of $L_{\beta+1}$ is the vocabulary of L_β plus:

[a] A 'semantic' predicate 'P_β (x)'.[12]

[b] A new conditional $\rightarrow_\beta$.

Limit:

The vocabulary of L_γ, γ a limit ordinal, is the union of the vocabulary of: $\{L_\beta : \beta < \gamma\}$

The general idea is that at each successor language we extend the expressive resources of our language by adding a predicate holding of (codes of) sentences that have

[10] In the formal theory we shall, for the sake of convenience, talk of an infinite series of languages. Strictly speaking, however, the view proposes, not infinitely many different languages, but a single language which is (or can be) successively extended, *ad transfinitum*. Reflecting this, each successive language in the formalism is a super-language of those that came before it.

[11] The falsity predicate is in fact redundant—falsity can be defined in the standard way in terms of truth and negation.

[12] For each $n > 1$, 'P_n(x)' is the nth 'pathogical' predicate.

the newest truth-value, and we also add a new conditional. Each conditional is a better approximation of the 'ideal' but unattainable conditional which satisfies all of the standard axioms and inference rules, including *modus ponens* and conditional proof.[13]

We assume that Gödel coding is carried out such that, for any sentence Φ in L_α, its Gödel code (written $<\Phi>$) relative to L_α, if it has one, is the same as its Gödel code relative to L_β for any $\beta > \alpha$. Note that there will be languages L_β such that some sentences of L_β will not receive Gödel codes (since if β is large enough, L_β is uncountable, yet we only have a countable infinity of numerals to serve as codes).

We construct a model for each language as follows: The model of L_0 is just any classical model of set theory $<D, I>$ where D is the domain and I is an interpretation function mapping sentences onto $\{t, f\}$. For each L_α ($\alpha \geq 0$) we construct an n+3-valued Kripke-style fixed point model $<D, I_\alpha>$ recursively as follows.

The truth-values (i.e. the range of I_α) are:
$\{t, f, n\} \cup \{p_\beta : \beta \leq \alpha\}$

where t is the value true, f is false, p_β is the β^{th} 'pathological' value, and n is a placeholder value given to sentences involving the application of a semantic predicate to a numeral that either is not the Gödel code of any sentence or is the Gödel code of a sentence which has not yet been added to the language (i.e. n is the value sentences get when they have not yet received a 'real' value). In other words, we do not assign a legitimate truth-value to the sentence '5 is true' (or its negation) if 5 is not the Gödel code of any sentence (in any of our languages), nor does the sentence '$5^{7568489}$ is true' obtain a legitimate truth-value in L_3 if the sentence coded by $5^{7568489}$ is a sentence which contains the predicate '$P_3(x)$'.

Given an assignment of semantic values to the atomic formulas in a language L_α, the interpretation of the logical connectives is determined as follows:

$$I(\Phi \wedge \Psi) = \min\{I(\Phi), I(\Psi)\} \text{ relative to the ordering:}[14]$$
$$n < p_\alpha < \ldots < p_{\beta+1} < p_\beta < \ldots < p_2 < p_1 < f < t$$
$$I(\Phi \vee \Psi) = \max\{I(\Phi), I(\Psi)\} \text{ relative to the ordering:}$$
$$f < t < p_1 < p_2 < \ldots < p_\beta < p_{\beta+1} < \ldots < p_\alpha < n$$

[13] The fact that one cannot have a wholly satisfactory conditional and a wholly satisfactory truth predicate in the same language is well known, and, like Field (2003), we opt here for retaining the naïve notion of truth, and settling for an account of the conditional (and thus the biconditional) which does not provide everything that we might wish for. In the final section of the chapter the present approach, which posits a never-ending series of conditionals that provide better and better approximations of the 'ideal' conditional, is contrasted with Field's approach, where a single, rather complicated, conditional is constructed in the hopes of getting the closest approximation possible.

[14] The clauses for the binary connectives are just a generalization of the weak Kleene scheme (Kleene (1952)). Interestingly, the strong Kleene scheme will not work in this framework.

$I(\neg\Phi) =$ f if $I(\Phi) = t$
t if $I(\Phi) = f$
$I(\Phi)$ otherwise.

For $\beta < \alpha$:

$I(\Phi \rightarrow_\beta \Psi) =$ n if $I(\Phi) = n$ or $I(\Psi) = n$

$\max\{I(\Phi), I(\Psi)\}$ relative to the ordering:
$f < t < p_1 < p_2 < \ldots < p_\beta < p_{\beta+1} < \ldots < p_\alpha < n$
if $I(\Phi) \neq n$ and $I(\Psi) \neq n$
and $I(\Phi) = p_\delta$ or $I(\Psi) = p_\delta$
where $\delta > \beta$

t if $I(\Phi) \neq n$ and $I(\Psi) \neq n$
and for any $\delta > \beta$, $I(\Phi) \neq p_\delta$ and $I(\Psi) \neq p_\delta$
and $I(\Phi) \leq I(\Psi)$ relative to the ordering:
$f < p_\alpha < \ldots < p_{\beta+1} < p_\beta < \ldots < p_2 < p_1 < t$

f otherwise.

The universal and existential quantifiers are treated as generalized versions of conjunction and disjunction respectively.

We now construct a sequence of models $<D, I_\alpha{}^\beta>$ via transfinite recursion (the final model $<D, I_\alpha>$ for each language L_α will be the least fixed point in this sequence). Given the clauses above, we need merely to specify the interpretation of atomic sentences in each model in the series.:

For any atomic sentence Φ of $L_{\alpha+1}$:

Base:

$I_{\alpha+1}{}^0(\Phi) =$ n if $\Phi = T(n)$ or $\Phi = F(n)$ or $\Phi = P_i(n)$ where n is not the Gödel code of a sentence in $L_{\alpha+1}$
$I_\alpha(\Phi)$ if Φ is a sentence of L_α.
$p_{\alpha+1}$ otherwise.

Successor:

If $\Phi = T(<\Psi>)$ then:

$I_{\alpha+1}{}^{\beta+1}(\Phi) = I_{\alpha+1}{}^\beta(\Psi)$

If $\Phi = F(<\Psi>)$ then:

$I_{\alpha+1}{}^{\beta+1}(\Phi) =$ t if $I_{\alpha+1}{}^\beta(\Psi) =$ f
$=$ f if $I_{\alpha+1}{}^\beta(\Psi) =$ t
$= I_{\alpha+1}{}^\beta(\Psi)$ otherwise

If $\Phi = P_\delta(<\Psi>)$ for some $\delta < \alpha$, then:

$I_{\alpha+1}{}^{\beta+1}(\Phi) =$ t if $I_{\alpha+1}{}^\beta(\Psi) = p_\delta$
$=$ f if $I_{\alpha+1}{}^\beta(\Psi) =$ t
$= I_{\alpha+1}{}^\beta(\Psi)$ otherwise

If $\Phi \neq T(<\Psi>)$ and $\Phi \neq F(<\Psi>)$ and $\Phi \neq P_\delta(<\Psi>)$ then:

$I_{\alpha+1}{}^{\beta+1}(\Phi) = I_{\alpha+1}{}^{\beta}(\Phi)$

Limit:

$I_{\alpha+1}{}^{\gamma}(\Phi) = \max\{I_{\alpha+1}{}^{\beta}(\Phi): \beta < \gamma\}$ relative to the partial ordering:
For any δ: $n < p_\gamma$, $p_\gamma < t$, $p_\gamma < f$, and $p_\gamma < p_\delta$ iff $\delta < \gamma$[15]

All operators, quantifiers, etc. in the above scheme are monotonic with respect to the partial ordering of truth-values utilized in the limit case above. Thus, the sequence of models described above will have a fixed point—that is—a model $<D, I_{\alpha+1}{}^{\beta}>$ such that:

$$<D, I_{\alpha+1}{}^{\beta}> = <D, I_{\alpha+1}{}^{\beta+1}>$$

$<D, I_{\alpha+1}>$, our model of $L_{\alpha+1}$, is just the minimal such fixed point (for the mathematical details of such fixed point constructions see Kripke (1972) or Fitting (1986)).

We extend the series of models into the transfinite by adding the additional clause handling atomic sentences in the base case for L_γ where γ a limit:

Base:

$I_\gamma{}^{0}(\Phi) = \max\{I_\beta(\Phi): \beta < \gamma\}$ relative to the partial ordering:
For any δ: $n < p_\gamma$, $p_\gamma < t$, $p_\gamma < f$, and $p_\gamma < p_\delta$ iff $\delta < \gamma$

Applying the same clauses as before for successor and limit superscripts, we are again guaranteed the existence of a fixed point. The minimal such fixed point is our model $<D, I_\gamma>$, of L_γ.

A few observations regarding the formal theory are in order (proofs have been omitted in order to keep this chapter concise and accessible):

[1] In any language L_α where $\alpha \geq 1$, $T(<\Phi>)$ and Φ are interchangeable *salve valutate* (i.e. the truth predicate is 'transparent', that is, $T(<\Phi>)$ and Φ always receive the same semantic value in $<D, I_\alpha>$).

[2] For any languages L_α and $L_{\alpha+\beta}$, and any sentence Φ in L_α (and thus in $L_{\alpha+\beta}$), $I_\alpha(\Phi) = I_{\alpha+\beta}(\Phi)$. In other words, once a sentence is expressible in one of our languages, and it receives a genuine truth-value (i.e. one other than n) in the model of that language, its truth-value does not change in models of extensions of that language.[16]

[15] In addition to guaranteeing the existence of a fixed point, the monotonicity of our valuation scheme relative to this ordering guarantees that the maximum referred to in this clause exists.

[16] A critic might be tempted to object to the present view on the ground that, as we pass from one language to another, the meaning of the truth predicate changes (which might, in turn, push us towards a view more like that sketched in Williamson (1998)). Fact [2] goes some ways towards assuaging such worries. The extension of the truth predicate does change from one language to the next, but only by adding new instances—no sentence can 'change' from true to some other value. Such changes in extension do not imply a change in meaning, however, any more than the production of the present volume changed the meaning of 'published anthology'.

[3] No model $< D, I_\alpha >$ makes true all instances of the T-schema:

$$T(<\Phi>) \leftrightarrow \Phi$$

on any definition of the biconditional.[17] Given [1], however, this is clearly not due to a fault in the truth predicate, but a failure to express a suitable conditional (and thus biconditional). The seriousness of this omission is lessened by the fact that each of languages makes true all instances of the T-schema for sentences of earlier languages. In other words, letting:

$$\Phi \leftrightarrow_\alpha \Psi \quad =_{df} \quad (\Phi \rightarrow_\alpha \Psi) \wedge (\Psi \rightarrow_\alpha \Phi)$$

$L_{\alpha+1}$ makes true all instances of the T-schema:

$$T(<\Phi>) \leftrightarrow_\alpha \Phi$$

where Φ is a sentence of L_α.

[4] Given [3], it is not surprising that we can express, within $L_{\alpha+1}$, the complete semantic theory for L_α.[18] For example, letting '$Sent_\alpha$ (x)' abbreviate the unary arithmetic predicate that holds of n if and only if n is the Gödel number of a sentence in L_α, and letting 'Conj(x, y, z) abbreviate the ternary arithmetic predicate expressing 'x is the code of the conjunction of the sentences that y and z code', we can, within L_3, formulate the clause for conjunction in L_2 as follows. First, we need to define slightly altered versions of our truth and pathological predicates:

$$\begin{array}{lll} T_{(\alpha)}(<\Phi>) & =_{df} & T(<\Phi>) \leftrightarrow_\alpha (T(<\Phi>) \leftrightarrow_\alpha T(<\Phi>))^{19} \\ F_{(\alpha)}(<\Phi>) & =_{df} & F(<\Phi>) \leftrightarrow_\alpha (T(<\Phi>) \leftrightarrow_\alpha T(<\Phi>)) \\ P_{(\alpha,\beta)}(<\Phi>) & =_{df} & P_\beta(<\Phi>) \leftrightarrow_\alpha (T(<\Phi>) \leftrightarrow_\alpha T(<\Phi>)) \end{array}$$

Intuitively, $P_{(\alpha,\beta)}(\Phi)$ gets the value t if Φ gets the α^{th} pathological value, gets the value p_δ if $\delta > \alpha$ and Φ gets the value p_δ, gets the value n if Φ get the value n, and gets f otherwise (similarly for $T_{(\alpha)}(<\Phi>)$ and $F_{(\alpha)}(<\Phi>)$). We can then formulate the clause for conjunction as:

$$\begin{array}{l} (\forall x)(\forall y)(\forall z)((Sent_2(x) \wedge Conj(x, y, z)) \rightarrow_2 \\ ((T_{(2)}(x) \leftrightarrow_2 (T_{(2)}(y) \wedge T_{(2)}(z))) \wedge \\ (P_{(2,2)}(x) \leftrightarrow_2 (P_{(2,2)}(y) \vee P_{(2,2)}(z))) \wedge \end{array}$$

[17] This is because none of our conditionals validates every instance of $(\Phi \rightarrow \Phi)$. For any sentence in L_α, however, $(\Phi \rightarrow_{\alpha+1} \Phi)$ is a theorem (in $L_{\alpha+1}$).

[18] In order to describe the semantics for languages L_α, where α is infinite, we make us of the fact that $L_{\alpha+1}$ contains a satisfaction predicate Sat(x, y) expressing the relation 'x is the code of a predicate satisfied by y' and a predicate $P^{th}(x,, y)$ expressing the relation 'x is the code of the yth pathological predicate'.

[19] $T_{(\alpha)}$ is the α-level 'strong' truth predicate, which, when applied to a sentence Φ, receives true if Φ is true, receives the $\alpha + 1^{th}$ pathological value if Φ does, and is false otherwise. Notice that these predicates, unlike the official truth predicate, are not transparent. If Λ is the simple Liar (and so Λ receives the first pathological value), then $T(<\Lambda>)$ also recieves the first pathological value, while $T_{(\alpha+1)}(<\Lambda>)$ is false.

$$(P_{(2,1)}(x) \leftrightarrow_2 ((P_{(2,1)}(y) \vee P_{(2,1)}(z)) \wedge (\neg P_{(2,2)}(y) \wedge \neg P_{(2,2)}(z))))\wedge$$
$$(F_{(2)}(x) \leftrightarrow_2 ((F_{(2)}(y) \wedge F_{(2)}(z)) \wedge (\neg P_{(2,1)}(y) \wedge \neg P_{(2,1)}(z) \wedge \neg P_{(2,2)}(y)$$
$$\wedge \neg P_{(2,2)}(z))))))$$

While the formal theory, as sketched, accurately models the philosophical picture as described in previous sections, there are a number of ways in which we could modify the details. Two are worth mentioning here.

The first way is to take, instead of the minimal fixed point, the maximal intrinsic fixed point (see Kripke (1972)). While studying the properties of various sorts of fixed points in this iterated version of Kripke's construction would no doubt provide us with a better understanding of the general framework as a whole, none of [1] through [4] above depend taking the minimal fixed point (we only need the weaker claim that we have taken some fixed point). So the choice between the minimal fixed point and the maximal intrinsic fixed point (or between these and some other fixed point) will depend on one's attitude towards ungrounded but determinate sentences such as:

D: D is either true or false.

The second way in which we might alter the present account is by adding more than one pathological truth-value at each extension of the language. Motivation for this idea is not hard to find. Recall that in the first section we saw that the Liar required a third truth-value because it fell into the category of sentences where 'what is says is the case if and only if what it says is not the case.' Given that truth and falsity are no longer being treated as exhaustive, this status is distinct from truth and falsity themselves. But it is not obvious that the status of the Liar, looked at in this way, is the same as that of other problematic ungrounded sentences. For example, the Truth Teller 'This sentence is true' falls into the 'what it says is the case if and only if what it says is the case' category, and the determinate sentence D above falls into the 'what it says is the case if and only if either what it says is the case or what it says is not the case'. Given that truth and falsity are not exhaustive (and thus, we cannot assume that, for every sentence, either what it says is the case, or what it says is not the case), it is not obvious that these two categories are identical to the status of the Liar sentence.[20]

Of course, the last paragraph is a bit rough and loose. Nevertheless, it does indicate that there might be good reasons for exploring the idea that extending our language by adding a new semantic predicate might introduce more than one new truth-value. Although adding such additional truth-values at each stage in the above construction

[20] In future work I intend on fleshing out this idea by treating the different truth-values as intimately connected to (and thus, within the semantics, representable by) the directed graphs associated with the different patterns of referential dependency exhibited by these different 'sorts' of pathological sentence. For an initial stab in this direction, one can consult Cook (2004).

greatly complicates an already complicated picture, the framework could be extended in this manner without greatly affecting the final shape of the account (in particular, versions of [1] through [4] will still hold). At this point, however, we will rest content with merely having pointed out the possibility of such extensions of the basic picture.

2.4 Dodging Revenge

On the account just sketched, the Revenge Problem is not a problem, but instead provides the crucial insight motivating the account: Given a language L, if we can completely describe the semantics of L, then we have (knowingly or not) extended our language to a new language L' (where such an extension involves not only adding to the set of wffs, but adding an additional truth-value that those wffs can receive). The semantic theory of L, however, as expressed in L' is not suffient for L' itself. In order to describe its semantics, we must extend the language again (and as a result extend the collection of truth-values as well). And so on. In hacker-speak, the Revenge Problem is no longer a bug—it is now a feature, exemplifying the indefinite extensibility of the concepts *language, statement*, and *truth-value*.

One advantage of this view is that there are no real limitations on what can be expressed. At any stage in the series of languages, we are free to extend the language we presently speak in order to describe all the facts (semantic and otherwise), although in doing so we might introduce additional truth-values (and thus enable ourselves to 'access' more facts). But further extensions will allow us to speak of those as well, and so on.

Of course, given this lack of expressive restriction, a critic might be forgiven for thinking that Revenge is likely to reappear. After all, if we can, at any stage in the game, extend our semantic resources in order to describe all of the (currently accessible) facts, then what is to stop us from extending the language so as to contain all possible semantic predicates at once?[21] In fact, doesn't the language used in the previous section, in describing the formal theory, amount to a language that does just that? But once we have allowed ourselves to extend the language in this way, there is nothing to stop us from forming the Super-Strengthened Liar:

SUP: This sentence is either false or has one of the pathological values.

[21] The comments of the next few paragraphs also explain why we cannot at any point add operators such as 'exclusion' negation to our language (i.e. a connective * such that *(P) is false if P is true, and true otherwise). Thanks go to Stewart Shapiro for pointing this out.

It is not difficult to see that such a sentence would be inconsistent in our formal theory, were it expressible. Fortunately for our account, however, there are good reasons for thinking that, contrary to appearances, SUP is not expressible in any language.[22]

Actually, there are two separate claims involved in this global version of the Revenge Problem which need to be distinguished, so each can be dealt with in turn. The first is that there is nothing to stop us from extending our language so that we can talk about all possible truth-values at once. The second claim is that we have already done this, in the previous section when formulating our formal account of the semantics.

Regarding the first claim, the initial answer is easy: there are reasons why we cannot formulate a language that contains every possible truth-value. Anytime we attempt to add a predicate 'is a pathological truth-value' to a language L_α, we end up extending the language, forming a new language $L_{\alpha+\beta}$. The semantics of $L_{\alpha+\beta}$ will require at least one more truth-value than that of L_α, and our new predicate will only be satisfied (i.e. receive the value t) by (Gödel codes of) sentences which receive values $p_1, p_2, \ldots p_\delta$ for some $\delta < \alpha + \beta$. This is the very lesson that the Revenge Problem teaches us: That any attempt to construct a language that allows us to talk about 'all' semantic values (in the sense of containing predicates for each value) brings a new pathological value into the picture, one which is not described by the language in question. In other words, the indefinite extensibility of the concept *statement* prevents us from every being able to say things about all statements (or, derivatively, about all truth-values).

At this point, however, we run into the second aspect of the objection. How can we claim that we can never talk about all truth-values at once, so the criticism goes, when we obviously quantified over all of them in the formal account given in the previous section? The easy answer to this question is that it misrepresents what exactly the formal model is doing. In particular, this objection confuses *describing* a language and *using* that very same language.

The formal semantics presented in §2.3 is a description (i.e. a model, in the intuitive sense of model) of a sequence of possible language extensions. No semantic predicates are used in describing this mathematical structure—the account is (or, can be reformulated) within first-order set theory. As a result, the formal model (can) occur in our (actual) base language corresponding to L_0 (and, in fact, this is the proper place

[22] Actually, this is a bit of a sloppy way of putting it, since we have expressed SUP, in English, just a few lines previously. More precisely, SUP is expressible (we just expressed it), but it does not say what we think it does. There are intricate questions looming here, connected to whether or not we are, in any sense, 'allowed' to extend our language in inconsistent ways. If, however, we assume that such a predicate as that used in SUP is added to our language, and that the language remains consistent, then the predicate does not express what we (in some sense) intended it to express. The situation is exactly analogous to the fact that we cannot add the term 'the set of all sets' to our language and expect it to consistently have its 'intended' meaning.

for such theorizing). Thus, the account of the formal semantics does not occur within a language that *uses* all of the semantic notions which it *describes* as occurring in the hierarchy.[23]

Nevertheless, the critic might continue, if our base language L_0 contains the semantic theory as described in the previous section, could we not extend the language by adding the predicate:

is a sentence which receives one of the pathological values *described* in L_0.

Furthermore, once we have added such a predicate to our language, what is to stop us from using this predicate to formulate a version of the Super Strengthened Liar?

The answer to this two-part question is of course a two-part answer. First, and most easily, we can grant to the critic that he is free to add such a predicate to his language (thus in a sense 'skipping' the step by step individual extensions of the language as described in the formal theory). The answer to the second question, however, is more complicated, and requires our getting a bit clearer on what we mean by the phrase 'pathological value described in L_0'.

Remember that L_0 (plus its interpretation) contains a set theory at least as strong as first-order ZFC. In addition, it describes a series of languages L_α for every ordinal α. Thus, there will be a least ordinal, call it π, such that there is a model of L_0 with π set theoretic ranks, and every model of π has at least π ranks. As a result, the theory of L_0 only guarantees the existence of all ordinals less than π (although it will have models with more ordinals as well). As a result, L_0 will only guarantee the existence of π distinct languages, with a corresponding collection of π distinct truth-values. Since a theory only 'describes' those objects that are guaranteed to exist according to that theory (or, at least, we assuming as much here), we should interpret the phrase 'pathological value described in L_0' as being satisfied only by pathological values p_β where $\beta < \pi$.

As a result, adding the suggested predicate to the language extends the language to a sub-language of L_π which requires the same truth-values for its interpretation as L_π itself (the language in question will be a sub-theory of L_π since the language we obtain—a natural language—is countable while L_π is not). L_π (and the new semantic value p_π) are not described in our (present) base theory.[24]

[23] Something similar to this approach is found in Field (2003). Although our accounts differ significantly, I owe much to the careful study of this chapter.

[24] Actually, it is this very phenomenon that justifies our carrying out the construction of the previous section into the transfinite. We did not do so because it is likely, or even possible, that we might some day master a language with an uncountably infinite vocabulary such as those high up in the hierarchy. Rather, the thought is that we might, through tricks such as the one considered here, extend our language in such a way as to require the same collection of truth-values as is required by one of the uncountable languages in the hierarchy (even if the actual language we are using remains countable).

So, given a particular set theory in the base language L_0, we can extend our language past all of the languages described in the formal semantics developed in L_0 by adding this predicate. This pushes us 'up' to a language beyond any described in our L_0 semantics. But of course we wish to be able to provide a semantics for this language as well.

The solution is to strengthen the set theory at the base level. If we add an additional axiom guaranteeing the existence of the ordinal π, then our new base theory will allow us to formulate a new account of the formal semantics which is identical to the original one other than the fact that it implies the existence of more languages (and more pathological values to accompany them). We can of course repeat the process, adding other semantic predicates, and obtaining even stronger languages as a result. In principle, there is no limit to such extensions (other than those imposed by our finite lifespans, etc.).

Thus, in a sense we can never provide a complete account of all of possible extensions of our base language, since for any such account (formulated in a particular set theory T), we can add a predicate:

is a sentence which receives one of the pathological values described in T.

which extends the language past what can be described by T. This is not a flaw in our formal semantics, however, but instead reflects a well-known feature of set theory. Our formal semantics entails that there is a language L_α for each ordinal α. But for any consistent set theory T, there is a stronger consistent set theory T' such that T' implies the existence of more ordinals than does T. Since we can never formulate a set theory which implies the existence of all possible ordinals, we can never formulate a formal semantics for our account which implies the existence of all possible extensions of our language (and corresponding truth-values) in some absolute sense of the word 'all'. While no single set theory implies the existence of all ordinals, however, there seems to be no reason to doubt that, for any ordinal, there is a set theory that implies its existence. As a result, for any possible extension of our language, we can formulate a semantics for it (by utilizing a suitably strong set theory in the base theory).

Earlier we drew an analogy between the indefinite extensibility of the concept *ordinal* and the indefinite extensibility of the concept *language* (and the corresponding indefinite extensibility of *truth-value*). The previous few paragraphs suggest, however, that there is more to this than just an analogy—in fact, the indefinite extensibility of our language just *is* the indefinite extensibility of the ordinals. This insight promises fruitful connections between the semantic and set-theoretic paradoxes.[25]

[25] One such connection involves the fact that certain very powerful versions of the Strengthened Liar, such as SUP above, seem (on the present account) to entail (at least indirectly) large cardinal axioms. These connections will be explored in future work.

Thus, the lesson to learn from the Revenge Problem is just this: What we can say (and the semantic values that what we say can receive) is indefinitely extensible in exactly the way the ordinals are. This implies that there is no language in which we can say everything. It does not imply, however, that there is something (coherent) which we cannot say in any language.

2.5 Truth, the Conditional, and Field

Something should be said, at this point, regarding the conditional, or, on the present view, the conditionals, plural. The formal theory presented in section 2.3 (and the informal account motivating it) provides, not a single conditional, but a new conditional for each extension of the language. The reason for this move, odd looking though it might be at first glance, is the need to avoid Curry's Paradox.

The arithmetic version of Curry's construction can be carried out as follows: Given an arbitrary sentence Φ, we can diagonalize to obtain a sentence X where:

$$X \leftrightarrow (T < X > \rightarrow \Phi)$$

If (a) our truth predicate is transparent (i.e. for any Ψ, Ψ and $T < \Psi >$ are intersubstitutable in all non-opaque contexts), (b) our conditional satisfies the inference rule modus ponens, and (c) our conditional validates the contraction axiom $((A \rightarrow (A \rightarrow B)) \rightarrow (A \rightarrow B))$, then we can prove Φ:

[1]	$X \rightarrow (T < X > \rightarrow \Phi)$	Diagonalization
[2]	$X \rightarrow (X \rightarrow \Phi)$	[1], Transparency of Truth
[3]	$(X \rightarrow (X \rightarrow \Phi)) \rightarrow (X \rightarrow \Phi)$	Contraction
[4]	$(X \rightarrow \Phi)$	[2], [3], Modus Ponens
[5]	$(T < X >) \rightarrow \Phi)$	[4], Transparency of Truth
[6]	X	[5], Diagonalization
[7]	Φ	[4], [6] Modus Ponens

Note that if we define negation in terms of the conditional and a primitive absurd sentence $\perp$:

$$\neg\Psi =_{df} \Psi \rightarrow \perp$$

then the Liar paradox is merely a special case of Curry's paradox.[26]

[26] The fact that the Liar Paradox can be seen as a special instance of Curry's Paradox, and thus in some sense the problems with semantic paradox in general hinge crucially on the conditional, is vastly underappreciated. Greg Restall's contribution to this volume, however, is a notable exception to this.

Thus, the real problem highlighted by the semantic paradoxes is that we cannot (on pain of triviality) have both:

(a) A transparent truth predicate.
(b) A conditional that uniformly satisfies the standard axioms and rules of inference for the conditional.

Since a transparent truth predicate is one of the main motivations for the present account, it follows that we cannot have a single conditional uniformly satisfying the rules usually attributed to the conditional. Denying the validity of basic principles such as contraction and modus ponens is a bitter pill to swallow, however, so the obvious move is to give up as 'little' of the standard rules for the conditional as possible. There seem to be two ways in which this can be accomplished.

The first way is to attempt to provide a single conditional that comes as close as possible to the standard classical account. Hartry Field's important (2003) (and its successors, including his contribution to this volume) represents such an approach. Field's theory contains a transparent truth predicate, and in addition it contains a conditional that is well behaved when applied to non-pathological sentences. The main drawback with this approach, however, is that the conditional cannot satisfy *all* instances of the standard axioms and rules for the conditional.[27] In particular, contraction can fail when the subformulas involved fail to receive classical values (as happens with the instance of contraction relevant to Curry's paradox).

On the other hand, however, we can give up on the idea of a single conditional, and instead accept that the concept *conditional* is itself indefinitely extensible. This is the approach taken above. In each extension of our language, we obtain (or, at least, we can obtain) a new conditional that is a better approximation to our intuitive ideas regarding 'if. . .then. . .'. Each conditional satisfies all the axioms and rules we would expect, at least when applied to sentences of earlier languages. In particular, every instance of *both* modus ponens and contraction are validated.

This last claim, of course, needs some explanation, given that if we have both modus ponens and contraction without any restrictions whatsoever, then we can (as we have already seen) reconstruct Curry's paradox. The point is this: all instances of the inference rule modus ponens are valid, since (assuming we define validity in the standard way as truth-preservation in models) no matter which instance of the conditional one takes, if it is true, and its antecedent is true, it follows that its consequent must be true. In addition, given any sentences A and B, there will be some

[27] This is not to say that his account does not do an admirable job, from a technical perspective, of validating as many of the standard rules as is possible.

ordinal α such that $((A \rightarrow (A \rightarrow B)) \rightarrow (A \rightarrow B))$ is a theorem.[28] Curry's paradox is blocked, however, since the conditional $\rightarrow_\alpha$ validating the relevant instance of contraction occurs later in the hierarchy than does the conditional occurring in the Curry sentence we obtain through diagonalization.

This would seem to be the main advantage of the present view over Field's approach, at least if one of our motivations (secondary, perhaps, to a transparent truth predicate) is to retain as much of the traditional account of the conditional as is possible. On the present account we need not give up any instance of any standard rule or axiom for the conditional (we just need to remember that not all instances of all rules or axioms are valid for all conditionals!).[29]

Despite this difference, both views (Field's and the present one) represent instances of a family of views which take as their primary motivation the salvation of a single, unified truth predicate that is transparent (at least in non-opaque contexts). Both views accept the fact that an account achieving this (and containing some

[28] Proof sketch: Let α be the least ordinal such that A and B both occur in L_α. Consider the $L_{\alpha+1}$ sentence:

$$((A \rightarrow_\alpha (A \rightarrow_\alpha B)) \rightarrow (A \rightarrow_\alpha B))$$

Since A and B receive, as truth-values, either t, f, or p_β for some $\beta \leq \alpha$, the relevant portion of the satisfaction clause for this conditional is:

$$I(\Phi \rightarrow_\alpha \Psi) \quad = \quad \begin{array}{ll} t & I(\Phi) \leq I(\Psi) \text{ relative to the ordering :} \\ & f < p_\alpha < \ldots < p_{\beta+1} < p_\beta < \ldots < p_2 < p_1 < t \\ f & \text{otherwise.} \end{array}$$

Assume, for *reductio*, that this formula fails to be true, that is:

$$I(A \rightarrow_\alpha (A \rightarrow_\alpha B)) > I(A \rightarrow_\alpha B)$$

Then:

$$I(A \rightarrow_\alpha (A \rightarrow_\alpha B)) = t$$

And:

$$I(A \rightarrow_\alpha B) = f$$

The former implies that:

$$I(A) \leq I(A \rightarrow_\alpha B)$$

And the latter implies:

$$I(B) < I(A)$$

Since partial orders are transitive, we obtain:

$$I(B) < I(A \rightarrow_\alpha B) = f$$

But this is impossible, since f is the minimal element of the ordering.

[29] In the end, however, the choice of one of these accounts over the other, or of some third view over both, should depend, not on technical merits, but on philosophical motivation.

sort of reasonable conditional) will require stratification of some sort (this is what Tarski had right), at least if we wish out final account to be able, in some sense, to 'completely' characterize all semantically problematic sentences. Field chooses to find the stratification in a hierarchy of stronger and stronger 'definite' truth predicates, while on the present view the same role is played by a hierarchy of stronger and stronger conditionals (plus the various semantic predicates that accompany them). Nevertheless, despite the vastly different details, the general philosophical viewpoint seems roughly the same.[30]

References

Cook, R. (2004). 'Patterns of Paradox', *Journal of Symbolic Logic* 69: 767–74

Dummett, M. (1993). *The Seas of Language*, Oxford,: Clarendon Press

Field, H. (2003). 'A revenge-immune solution to the semantic paradoxes', *Journal of Philosophical Logic* 32: 132–177

Fitting, M. (1986). 'Notes on the mathematical aspects of Kripke's theory of truth', *Notre Dame Journal of Formal Logic* 27: 75–88

Godel, K., (1992). *On Formally Undecidable Propositions.* New York,: Dover

Kleene, S. (1952). *Introduction to Metamathematics*, Amsterdam,: North Holland

Kripke, S. (1975). 'Outline of a theory of truth', *Journal of Philosophy* 72 (1975), 690-716; reprinted in Robert L. Martin (ed.), *Recent Essays on Truth and the Liar Paradox*, Oxford, Clarendon Press (1984), pp. 53–81

Russell, B. (1906). 'On some difficulties in the theory of transfinite numbers and order types', *Proceedings of the London Mathematical Society* 4, 29–53

Shapiro, S., and Wright, C. (2006). 'All things indefinitely extensible', in A. Rayo and G. Uzquiano (eds.), *Absolute Generality*, Oxford: Oxford University Press (2006)

Tarski, A. (1933). 'The concept of truth in the languages of the deductive sciences', *Prace Towarzystwa Naukowego Warszawskiego, Wydzial III Nauk Matematyczno-Fizycznych* 34, Warsaw; English translation in Alfred Tarski, *Logic, Semantics, Metamathematics, Papers from 1923 to 1938*, John Corcoran (ed.), Indianapolis,: Hackett Publishing Company, (1983), pp. 152–278

Williamson, T. (1998). 'Indefinite extensibility', *Grazer Philosophische Studien* 55: 1–24

[30] Earlier versions of this chapter were presented at *The Ohio State University and Arche: The AHRC Centre for the Philosophy of Logic, Language, Mathematics, and Mind*, and the current version has benefited much from the comments, criticisms, and congenial atmosphere found there. Thanks are also due to JC Beall, Stewart Shapiro, Timothy Williamson, and an anonymous referee for additional comments or guidance. A special debt is owed to the students in my Spring 2006 undergraduate Leibniz seminar at *Villanova University*, who were forced to listen to early versions of these ideas before and after class, and without whose prodding I might never have started, or finished, the chapter.

3

The Liar Paradox, Expressibility, Possible Languages

Matti Eklund

3.1 Introduction

Here is the liar paradox. We have a sentence, (L), which somehow says of itself that it is false. Suppose (L) is true. Then things are as (L) says they are. (For it would appear to be a mere platitude that if a sentence is true, then things are as the sentence says they are.) (L) says that (L) is false. So, (L) is false. Since the supposition that (L) is true leads to contradiction, we can assert that (L) is false. But since this is just what (L) says, (L) is then true. (For it would appear to be a mere platitude that if things are as a given sentence says they are, the sentence is true.) So (L) is true. So (L) is both true and false. Contradiction.

In the literature there is a bewildering variety of purported solutions to the liar paradox. I will not discuss any of these purported solutions in any detail. Instead I will further problematize the question of what a solution should achieve. I will bring up a somewhat neglected cluster of problems connected with the liar paradox. These problems remain even if one of the solutions to the liar paradox currently on offer would succeed perfectly in solving the problem it is designed to solve. They arise when we consider *what possible languages there are*.[1] Often in discussions of the liar paradox, truth

[1] (1) I will throughout conceive of languages as necessarily existing abstract objects. If this assumption should be rejected, my conclusions about this will just have to be reformulated as claims about *what*

predicates are treated as if they were applicable only to sentences (in contexts), and then only to sentences of one language. Clearly this is a simplification. The English language predicate 'true' can be applied to sentences of all sorts of languages, to utterances of all sorts of languages, and, perhaps primarily, to propositions. As the arguments here will show, the simplification is not innocent, but is an important distortion.

I will approach the problem of possible languages via an issue that has attracted some attention in the liar literature: that of the liar paradox and expressibility. As is well known, a consequence of many of the most popular and most widely discussed solutions to the liar paradox is that some properties are deemed inexpressible in natural language. Theorists disagree about just which notions, if any, are inexpressible in natural language; how problematic or not such consequences are; and how best to conceive of the inexpressibility.

In section 3.2, I will outline the main views on the liar paradox and expressibility. Sections 3.3 and 3.4 will be devoted to critical discussion of the most important of these views. Section 3.5 will focus on the issue of what possible languages there are. In section 3.6, I will discuss how serious revenge problems really are.

3.2 Inexpressibility and Revenge Problems

Whatever is in the end the correct account of the liar paradox, the liar reasoning undoubtedly *establishes* what is sometimes called *Tarski's theorem*: in no language whose logic is classical and which can talk about itself to a sufficient extent can there be a predicate that *satisfies the T-schema*, where for a predicate to satisfy the T-schema is for a valid schema to result when this predicate is substituted for the 'T' in

s is T iff p,

(where instances of this schema are obtained by putting sentences for p and names of the corresponding sentences for s).

A natural thought is that Tarski's theorem entails that natural language is expressively limited. The reasoning would be that natural language satisfies the

there could be. (2) There are independent problems concerning the notion of all languages. Focus, more narrowly on a problem concerning talk of all possible predicates. (If there is a problem concerning all possible predicates, there is a problem concerning all possible languages.) If, no matter what objects the xs are, there is a predicate true of exactly the xs, there are more possible predicates than there are objects—a contradiction, if predicates are objects. We must either think of quantification over predicates and languages as restricted, or else take the space of languages and predicates as smaller. Serious though these problems are, I will set them aside as orthogonal to the problems that I will mainly focus on.

conditions stated in Tarski's theorem: so the property which satisfies the T-schema is not expressible in natural language, where a property satisfies the T-schema iff it is such that any predicate which expresses it satisfies the T-schema. (Notice that here and throughout I will use 'expresses' for the relation between predicates and the corresponding *properties*.) There are two ways to resist this reasoning. One is to deny that natural language does satisfy the conditions of Tarski's theorem (the relevant type of sentence cannot be meaningfully formed or the logic of natural language is not classical).[2] The other is to deny that there is a property which satisfies the T-schema. The thought in the latter case would be that the property simply doesn't exist. However, if there can be a richer metalanguage with a predicate which behaves in the right way semantically—specifically, which has the right semantic features to express the property—then it is hard to maintain that the property does not exist. (Here, and throughout, I will assume an 'abundant' ontology of properties.)[3]

Someone who takes Tarski's theorem to show that there is a property which cannot be expressed in natural language might further hold that any predicate expressing the property of being true would have to satisfy the T-schema: so what would be established is that truth cannot be expressed in natural language.[4] Sometimes, e.g. in Robert Martin's (1984a), this reasoning is attributed to Tarski himself. This suggested conclusion is on the face of it bizarre. The claim is that truth cannot be expressed in natural language. But is truth not expressed by the predicate 'true', undoubtedly a natural language predicate?

The threat that the liar reasoning entails that natural language is expressively limited comes up even if we deny that natural language satisfies the conditions of Tarski's theorem, as in fact many theorists writing about the liar do. Most popular theories of the liar paradox seem to have the conclusion that certain properties—including properties which are expressed by predicates employed in these theories—cannot, on pain of consistency, be expressed in natural language but only in a richer metalanguage.

Take first theories of the liar which attempt to solve the paradox by saying that the liar sentence has a semantic status somehow intermediate between truth and falsity. The most famous theory of this kind is Saul Kripke's (1975). Sometimes this intermediate status is conceived of as a third truth-value; sometimes, as in Kripke, it is rather conceived of as 'unsettled', or the absence of a truth-value. I will call sentences with this intermediate status *neuter*; and I will for simplicity talk about this

[2] There are non-trivial problems concerning going back and forth between what holds for formal theories and natural languages.

[3] Below I will present a justification of this choice and explain why it does not beg any important questions.

[4] Or, focusing on language-specific truth predicates as the literature does, that truth in L cannot be expressed in L, for L a natural language.

as a truth-value, even if this begs otherwise important questions concerning how the intermediate status is best conceived. A liar sentence of the kind we started talking about can consistently be said to be neuter. But a purported solution of this general kind immediately invites the *strengthened liar*. Consider a new liar sentence which says of itself that it is *untrue*. A sentence that is neuter would appear to be, among other things, untrue. But then we are obviously back in paradox. There are different ways around this problem. But one popular way out is to say that in a three-valued setting, both 'not' and 'true' are 'weak': that they take neuter into neuter.[5] If a sentence S is neuter, so is its negation and so is a sentence which says of S that it is true. If the negation sign and the object language truth predicate work this way, then a sentence which says of itself that it is not true is after all not paradoxical: it can consistently be ascribed the value neuter. Note it is not enough for paradox to be avoided that 'not' and 'true' be in this way weak: so long as there is any construction at all in the object language that takes both neuter and false into truth and takes truth into falsity—so long as *strong negation*, as it is often called, can be expressed at all—we land in paradox. The position must be that no construction in the object language can express that a sentence has some semantic status other than truth: as I will put it, that a sentence is *untrue*. Kripke also holds that untruth can be expressed only in a richer metalanguage.

Second, consider the revision theory of truth, defended most prominently in Anil Gupta and Nuel Belnap (1993).[6] The revison theorist retains classical logic but holds that sentences like the liar sentence cannot *stably* be assigned any truth-value. This theorist cannot, it seems, allow that 'not stably true' of the object language expresses what we would naïvely take this predicate to express. For then consider a liar sentence that says of itself that it is not stably true. Like Kripke, Gupta and Belnap hold that some semantic notions needed in a semantic theory for a natural language can be expressed only in a richer metalanguage: that natural language is not *semantically self-sufficient*. Gupta (1997) argues at some length that the ideal of avoiding this consequence may well be unattainable.[7]

In general, the situation is this. A standard kind of *revenge problem* for purported solutions to the liar paradox is the problem that given the expressive resources used to solve the solution to the liar in its simple form, a new paradox can be formed. The standard form of the revenge problem is this: the expressive resources of our language allow us to exhaustively and exclusively divide sentences into the true ones and the rest. If our language has sufficient expressive resources to state

[5] See e.g. Kripke (1975), Soames (1999), and Maudlin (2004).

[6] Earlier presentations of the revision theory are Belnap (1982), Gupta (1982), and Herzberger (1982 and 1982a).

[7] Gupta (1997), pp. 439 ff.

an exhaustive and exclusive division of all sentences into the true ones and *the rest*, paradox can be reinstated. Just let our new liar sentence say of itself that it belongs to the rest.[8]

One common way to avoid a threatening revenge problem is to deny semantic self-sufficiency. The revenge problem arises only if it is assumed that the semantic theory can be formulated in the object language. It is only then that a new liar sentence can be formulated in the object language. But it is often regarded as an embarrassment for a theory of the liar paradox if it is forced to, so to speak, push some predicates into such a metalanguage. That a theory is so forced can often seem suspect already on intuitive grounds: it does not seem as if I expand my language when I use a construction expressing strong negation or when I use an expression expressing the property of not being stably true. There are also some arguments in the literature to the effect that any theory which is so forced simply must be unacceptable: natural language must be regarded as semantically self-sufficient. One argument is suggested by Vann McGee (1994 and 1997): human language lies within the reach of human understanding and hence it must be possible to state a correct theory of human language in a human language.[9] However, even if we agree on the assumption here, the conclusion that we must be able to give a semantic theory for a natural language L in L does not follow. Whereas the assumption might entail that there is some possible natural language where we can give a semantics for English, and generally that for every natural language there is some natural language where we can give a semantic theory for it, this does not mean that *English* possesses those resources. The assumption at most entails that for every natural language L there is some language we are capable of employing in which a semantics for L can be given.

Graham Priest (1990) presents a different argument for semantic self-sufficiency. He says, 'The whole point of *solutions* to the liar paradox (as opposed to reformist suggestions as to how to change our language) is to show that our semantic discourse (about truth, etc.) is, appearances notwithstanding, consistent. An attempt to show this which produces more and more discourse, not in its own scope, therefore fails.'[10] This is not a convincing argument either. First, even if Priest is entirely right, one might think that the proper conclusion to draw might be that a solution of the kind described is unattainable. (Compare again Gupta's stance.) Second, there would be a problem if the further discourse 'produced' could not be plausibly believed to be

[8] Priest (1987), p. 29, makes this point well.

[9] See McGee (1994), p. 628, and (1997), p. 402. f; compare too Gupta (1997), pp. 440. ff, who criticizes the argument. McGee comes closest to actually endorsing the argument discussed in the main text in his (1994); the discussion in his (1997) is considerably more guarded.

[10] Priest (1990), p. 202.

consistent. But there is no immediate reason why the hierarchy of metalanguages that the Kripkean is saddled with should fail to be consistent.[11, 12]

In the above discussion I have, inter alia, suggested a number of different views on the implications of the liar paradox regarding the expressive limitations of natural languages. Let me now be more systematic:

RADICAL INEXPRESSIBILITY (RI). Some seeming ordinary semantic properties, like truth, are in fact not expressible in ordinary natural languages, but only in a richer metalanguage.
INEXPRESSIBILITY (I). Some semantic properties—including semantic properties we need to be able to express in an adequate semantic theory of natural language—are expressible only in a metalanguage.
SEMANTIC SELF-SUFFICIENCY (SS). All semantic properties we need to be able to express in an adequate semantic theory of natural language are expressible already in natural language.
WEAK UNIVERSALITY (WU). All properties are expressible in natural language; or at any rate: the liar paradox casts no doubt on this claim. (It is the qualification that warrants the 'weak'.)

(RI) entails (I) but not vice versa. (WU) entails (SS) but not vice versa. Martin ascribes view (RI) to Tarski. View (I) has the best claim to being orthodoxy: it is subscribed to by both Kripke (1975) and Gupta and Belnap (1993). McGee's and Priest's arguments are arguments for (SS). McGee is explicitly doubtful about (WU). Although Priest does not outright endorse (WU), some of the moves he makes are explicable only on the assumption that he holds this stronger view. For example, having shown to his satisfaction that Boolean negation (or, to take this in terms of properties, the property that a sentence has when its Boolean negation is true) need not be expressed in a semantic theory for natural language, Priest anyway sees it as incumbent upon himself to provide an argument for why Boolean negation just is not there to be expressed, and that it is not an expressive limitation of English that English does not possess the resources to express it. Hartry Field should also be mentioned along with McGee and Priest as someone defending a view of type (SS) or (WU). Field's theory of truth is designed to meet the requirement of self-sufficiency, and Field takes this to be a supremely important consideration in its favor.[13]

[11] It is of course unclear what it is for a 'discourse' or a 'language' to be consistent or inconsistent. But this unclarity is a problem for Priest, not for me.

[12] Similar remarks as apply to McGee and to Priest seem to me to apply to the argument for self-sufficiency given in Scharp (manuscript). Scharp gives an argument for why a theory of truth for a language needs to be *internalizable*: roughly, that even if it cannot be given in our actual natural language L there must be some expanded version of natural language, L+, where it can be given and such that the theory also applies to L+. The argument is that if this internalizability requirement is not met the theory cannot be both 'descriptively correct' and 'descriptively complete'. Even if Scharp is right about this, the conclusion may well be that a semantic theory of the kind described simply cannot be had.

[13] See Field (2003a), (2003b), (2005a), (2005b), (this volume).

I will in my discussion focus primarily on (I) and (WU). Given that the arguments for semantic self-sufficiency are unpersuasive, I fail to see (SS) as a principled position. As for (RI), I happen to think that a brief dismissal of it as absurd is too quick—more on this at the very end of the Chapter—but I will anyway not focus much on it. The problems I will discuss regarding the more popular view (I) apply in a similar way to (RI). In section 3.3, I will discuss (I). In section 3.4, I will discuss the viability of (WU).

3.3 Universality

I will approach (I) via consideration of the question of the *universality* of natural language.[14] The notion of universality was introduced by Tarski:

> A characteristic feature of colloquial language . . . 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'.[15]

What does 'universality' mean exactly? Instead of considering what Tarski and other theorists might have meant by it let me introduce my own—admittedly rough—characterization (which however seems to be in the spirit of Tarski):

> A language L is universal iff for every property there is a predicate of L that expresses it.[16]

A few remarks on this characterization are in order: *First*, for simplicity I talk only about the expressibility of *properties*. In a properly general characterization of the expressive power of natural languages we should consider not only properties, but also logical operations, objects, etc. *Second*, given the focus on properties, metaphysical questions concerning the nature and abundance of properties become relevant. On a conception of properties given which there are only very few properties (a sufficiently 'sparse' conception of properties, in common jargon) the claim of universality can be (relatively) trivial, and not the interesting claim it is intended to be. (Consider for instance a view on which the only properties there are, are properties corresponding

[14] The classic discussion of universality is Tarski (1935/83); see also Fitch (1946 and 1964), Gupta (1997), Herzberger (1970), Martin (1976), McGee (1991 and 1997), Priest (1984 and 1987), and Simmons (1993).

[15] Tarski (1935/83), p. 164. It is because Tarski says these things that (RI) cannot unproblematically be attributed to him.

[16] The characterization of universality employs unrestricted quantification over properties. Such quantification is obviously problematic. Although the problems concerning the possibility of unrestricted quantification over properties are not unrelated to the liar paradox, I will not here attempt to directly address those problems.

to the predicates employed in microphysics.) As already mentioned, I will throughout assume an abundant conception of properties, given which, roughly speaking, there are maximally many properties. A justification for this strategy is that the talk of properties can be regarded as a convenient way of talking about what objects fall under a predicate. *Third*, there are questions about how to individuate languages. In logic, languages are individuated very finely: add another constant and you have a new language. Outside logic, however, natural languages are individuated more liberally: a language can undergo significant changes in its vocabulary and its syntax and still we happily speak of it as the same language after these changes. I will here individuate natural languages finely. The justification for this is straightforward. For 'English is universal' to express the substantive claim it is meant to express, it must not be sufficient for its truth that for every property, English as it currently exists could be expanded with a predicate which expresses this property, while still we would call it the same language. *Fourth*, there are some questions regarding just how to understand 'expressed'. These questions I will soon get to.

Now, despite what unclarities may remain in the given characterization of universality, the characterization ought to be clear enough that it should seem intuitively implausible that all natural languages are universal. Consider, say, eighteenth-century English. Do we really want to say that eighteenth-century English contained the resources for expressing the property of being a quark, or the property of being an inaccessible cardinal? Moreover, is it not perfectly possible that there are properties that simply are determinately outside our cognitive reach, in the way that (I would suppose) the property of being an inaccessible cardinal is outside the cognitive reach of a gorilla?

But there is a complication. Stick with the example of eighteenth-century English. Intuitively, it is in some respects limited in expressive resources. It cannot express the property of being an inaccessible cardinal. So there is some property that can be expressed in some natural language that cannot be expressed in eighteenth-century English. But suppose the tallest man on Earth at midnight 1 July 2016 employs the predicate 'is an inaccessible cardinal'. Then consider the predicate of eighteenth-century English 'falls under the predicate employed by the tallest man on Earth at midnight 1 July 2016'. This predicate is true of exactly the inaccessible cardinals: and so, arguably, expresses the property of being an inaccessible cardinal. Hence it appears that our verdict on eighteenth-century English must be revised: it is after all universal.[17]

There is an immediate objection to what has just been argued. The objection is that what is shown is at most that for every property ϕ there is a predicate of

[17] Objection: did eighteenth-century English really contain linguistic vocabulary like 'falls under' and 'predicate'? Reply: The details are irrelevant. What I take for granted is that eighteenth-century English contained some linguistic vocabulary that could be employed for the same purpose.

eighteenth-century English which, given that the circumstances are propitious, as used on a particular occasion is true of exactly what has ϕ. But this is not enough for the predicate to express ϕ. The predicate 'falls under the predicate employed by the tallest man on Earth at midnight 1 July 2016' does not express untruth, even if the predicate the tallest man on Earth then used did express this property; rather, it expresses the property of falling under the predicate employed by the tallest man on Earth at midnight 1 July 2016. That is different.

However, while there is a distinction here, it is actually for present purposes possible to disregard it. Let us say that a property ϕ is *indirectly expressible* in a language L iff there is a predicate F of L such that for some context c, an utterance of F by a speaker of L is such that for all x, 'F(x)' and 'x has ϕ' have the same truth-value. It would be wrong, for the reason just noted, to identify indirect expressibility as expressibility.[18] But all that the present discussion needs is indirect expressibility of the relevant properties. The reason is that provided that given that a certain theory of truth demands that property not be expressible in English, it also demands that this property not be indirectly expressible in English.

Let me take a few more examples to illustrate further what is at issue, before returning to the lessons of the liar paradox. Suppose, as earlier suggested, that there are properties beyond our cognitive reach: they are too complex, too fantastic, too divine, what have you. These too are properties we can express in our ordinary language, if what I have suggested with respect to the previous examples is correct. Consider predicates of the form 'falls under the predicate employed at t by the most F creature in the universe' (Where F might be *intelligent*, or *old*, or *morally praiseworthy*, or what have you.) Some of these predicates might be true of exactly what has one of these properties which supposedly are beyond our ken. It is unlikely, perhaps, that for every cognitively unreachable property (as we might call these properties) there will in fact be a predicate of our language which is in fact true of whatever has that property, as things stand. (Some cognitively unreachable properties need not be thought of at any time by any creature in the universe, etc.) But for all that, each of these properties may still be indirectly expressible.

Compare too an objection Keith Simmons presses against taking natural languages to be universal: for every set in the ZF hierarchy there is a distinct concept (the concept of being a member of that set), and '[g]iven certain assumptions about natural languages (in particular, about upper limits on the size of vocabularies and the length of sentences), these concepts would outrun the expressive capacity of any natural

[18] Compare the notion of 'loose speaker expressibility' characterized in Hofweber (2006): a property ϕ is loosely speaker expressible in L iff there is a predicate F of L and a context C such that that an utterance of F (by a speaker of L) in C expresses ϕ.

language'.[19] This is a convincing argument against the possibility of the *expressibility* of all properties in natural language. However, Simmons' reasoning leaves open that these concepts should all be *indirectly* expressible in natural language.[20]

Return now to the liar paradox. The present considerations can appear to have the consequence that any theory of the liar paradox that relies on kicking some predicates up into the metalanguage can be easily refuted. Let U be the relevant metalanguage predicate. Then in the object language we will have predicates of the form 'falls under the predicate U of the language spoken by so-and-so at time so-and-so'. Whatever is expressed by U will be indirectly expressible in the object language. But then we are back in paradox: let P, in a given context, indirectly express what U expresses, and consider a sentence which says of itself that it is P. (Call this kind of version of the liar paradox an *interlanguage paradox*.)

However, it would be much too hasty to conclude from this reasoning that no view of type (I) can be maintained. For the reasoning relies on the assumption that for any predicate F, F and 'falls under F' are coextensive—as I will put it, that 'falls under' is *transparent* (generally, that satisfaction predicates like 'falls under', 'is true of', 'satisfies', etc. are transparent). While it is natural to hold that 'falls under' is transparent, liar-like reasoning—specifically, Grelling's paradox[21]—calls this into doubt. More specifically, the assumption that 'falls under' is transparent is in the same category as the assumption that a sentence S and the corresponding sentence 'S is true' always have the same truth-value (that 'true' is transparent). Both assumptions are called into doubt by liar reasoning.

But although there is no immediate refutation of views of type (I) in the offing, there is a nearby puzzle that deserves stressing. Although some take the liar reasoning to show that the truth and satisfaction predicates of English are not transparent, for example Kripke (1975) and others following up on his work have shown that there can be languages which are like natural languages whose truth and satisfaction predicates are transparent. One way the truth and satisfaction predicates of a language L can be transparent is if L lacks an untruth predicate. This is for instance the way that Kripkean theories of truth achieve transparency. The languages he describes

[19] Simmons (1993), p. 15. Simmons talks about the expressibility of *concepts*; I prefer to think of the issue as about the expressibility of *properties*.

[20] Once the notion of indirect expressibility is introduced, one can introduce a corresponding notion of *indirect universality*: a language L is indirectly universal iff for every property there is a predicate of L that indirectly expresses it. What I have argued is that some immediate reasons for doubt about whether English is universal do not readily show that English fails to be indirectly universal. There may, for all I am concerned to argue, be reasons to doubt whether English really is indirectly universal. All I am concerned with is the indirect expressibility of those properties some theorists say for liar-related reasons can only be expressed in a metalanguage.

[21] Grelling's paradox: Let a predicate be *heterological* just in case it is not true of itself. Is 'heterological' true of itself or not?

can contain transparent truth and satisfaction predicates on pain of not containing untruth predicates, as earlier discussed. If they did, we would be back in paradox, as is well known. But here we must be careful. Can there be a language L which contains transparent truth and satisfaction predicates on pain of not containing an untruth predicate?[22] There is a difficulty here. There cannot be such a language L, if (i) there is some other language L* with an untruth predicate for languages of the kind to which L, if it exists, belongs (say, natural languages), and (ii) L's truth and satisfaction predicates apply to sentences and predicates of languages other than L; specifically, of L* (if it exists). Now, if we think of L as something like a natural language, L would appear to have to satisfy (ii). It then comes down to condition (i). We are in a curious position. For L can exist if L* doesn't and L* can exist if L doesn't. Which of these two languages exists?

I will return below to the significance of this kind of puzzle.[23] For now let me just note the following. The puzzle does not have to do with what we should say about actual natural languages. The question is about a hypothetical natural language L, like actual ones, except for the possible difference that its truth and satisfaction predicates are transparent and on pain of consistency it does not contain an untruth predicate. One might well think that L, whether actually used or not, certainly exists. But now it turns out that L does not exist, if there is another language L*, with an untruth predicate for languages like L.

The sort of puzzle I have called attention to does not refute view (I) on semantic self-sufficiency and universality. All it shows is that such a view can be hostage to facts about what possible languages there are.

3.4 Priest on Boolean Negation

Turn now to views of type (WU): views according to which the liar paradox does not have the consequence that some notions are inexpressible. The discussion will be

[22] Note the formulation. The question is not about the possibility of transparent truth and satisfaction predicates, simpliciter. It is about one particular way of achieving transparency.

[23] It deserves noting that while the puzzle presents problems for Kripkean theories of truth, it presents no problem for more old-fashioned, Tarskian theories, denying that truth in L can be expressed in L. Two points deserve noting. (i) A consistent Tarskian ought to say the same thing about 'true of', 'satisfies', and 'falls under' as about 'true'. (ii) A Tarskian obviously wouldn't hold that L can express truth in other languages. (After all, truth in a natural language can for the Tarskian be expressed in a metalanguage. But a Tarskian will obviously want to hold that there cannot be pairs of languages L and L′ such that L can express truth in L′ and L′ can express truth in L.

It should be clear that no interlanguage paradox can force the Tarskian to accept that truth and satisfaction in L can at least be indirectly expressed in L.

focused on Graham Priest's defense of a view of this type. Priest is a dialetheist, and this informs his discussion. But even though dialetheism is a radical view, it should be clear that the points I make regarding Priest generalize.

A dialetheist holds that there are true contradictions. The liar paradox presents one of the main arguments for dialetheism. The idea is that the solution to the liar paradox is to recognize true contradictions (the liar sentence is both true and not true). For the view that there are true contradictions to get off the ground, negation must not work in such a way that from a contradiction everything follows (it must not satisfy *ex falso quodlibet*); negation must, in other words, be *paraconsistent*. For otherwise the dialetheist would have to hold that all propositions are true.

One of Priest's main arguments for a dialetheist solution to the liar paradox is this. The liar reasoning forces us to make a choice between embracing expressive incompleteness (hence to adopt view (RI) or (I)) or embracing inconsistency. Non-dialetheist theories of the liar commit their proponents to embracing the former alternative. But that is untenable. Hence the latter alternative, dialetheism, is forced upon us.[24]

There are a few different lines of resistance to an argument like this. One is to deny that non-dialetheic theories have the consequence that natural languages are expressively incomplete. A second is to deny that this consequence is particularly damaging. A third, which I will focus on in this section, is to argue that the dialetheism does not in fact avoid expressive incompleteness. Such an argument can take different forms. It can be argued that the dialetheist cannot allow a predicate which expresses that a sentence is *only* untrue. For consider then a liar sentence which says of itself that it is only untrue. The standard dialetheist strategy, of saying that liar sentences are both true and untrue, does not seem to be applicable here. For if this sentence is both true and untrue, then it must be only untrue. But if it is only untrue, then it is true after all.[25] Another argument, which is the one I will focus on, since it is the one that Priest (1990) is about, is to the effect that the dialetheist cannot allow the expressibility of *Boolean* negation. As Priest himself states the problem, 'If [Boolean negation] is allowed then, using the T-scheme and self-reference in the usual way, we can produce a sentence equivalent to its own Boolean negation, and hence deduce a Boolean contradiction, whence everything follows by Boolean EFQ [*ex falso quodlibet*]'.[26] Talk about the inexpressibility of Boolean negation is on its face different from talk about inexpressibility of properties, since a negation sign does not express a *property*.

[24] See e.g. Priest (1990), p. 202.

[25] The point of the argument (which has been presented by others, see e.g. Parsons (1990)) is just that dialetheism faces its own expressibility problem, prima facie as serious as the expressibility problems that arise for other theories. Naturally, dialetheists have responses to these problems; see e.g. Priest (1995).

[26] Priest (1990), p. 203.

But I will slide over this difference, since the same remarks that apply to the question of the possibility of having a logical operator that expresses Boolean negation apply to the question of the possibility of having a predicate that expresses the corresponding property (the property a sentence has when its Boolean negation is true).

Priest's way out is to deny that there is such a thing as Boolean negation. It just is not there to be expressed. So the supposed fact that no expression of English expresses it indicates no expressive limitation in English. His argument for this perhaps surprising conclusion proceeds by way of criticizing possible attempts to show that there must be such a thing as Boolean negation. Let me briefly summarize how Priest argues. [27]

Suppose first that someone says that we can state rules of inference characterizing Boolean negation, and therefore Boolean negation exists. To this the reply is that we can state rules of inference for Prior's 'tonk'—the disjunction introduction rule and the conjunction elimination rule—and this does not mean that there is an associated operation expressed by the connective.[28] A more sophisticated version of the appeal to rules of inference would have it that the rules of inference for Boolean negation satisfy the condition for successful introduction of a connective while those for 'tonk' do not. But a standard condition is *conservativeness*: and a connective satisfying the rules of inference for Boolean negation will not satisfy this condition, if it is introduced into a dialetheist language with a truth predicate satisfying the T-schema (precisely because it makes the language trivial).[29] A different suggestion is to say that the rules of inference manage to characterize an operation if they are demonstrably sound. Concerning this, Priest argues that any argument to the effect that the relevant rules of inference are sound must itself employ Boolean negation, and so is question-begging.[30] As Priest is quite clear about, the force of his argument against the existence of Boolean negation relies on there being an independent case for thinking that dialetheism is true of English.[31] Consider the point about how the rules of inference for Boolean negation fail to satisfy the conservativeness criterion. That point relies on assumptions about what is in the language prior to the introduction of Boolean negation. Priest argues that this fact about the dialectic is no cause for concern.

[27] My discussion of this argument will follow Priest (1990). The argument is also given in Priest (2005), ch. 5, which however focuses less on the liar paradox.

Another type of argument for why some things we thought were expressible at least in some possible language do not in fact exist is that given in Maudlin (2004), passim. Maudlin goes from the assumption that 'truth and falsity are always rooted in the state of the world' to the claim that 'if a sentence is true or false, then either it is a boundary sentence, made true or false by the world of non-semantic facts, or it is semantically connected to at least one boundary sentence, from which its truth value can be traced', (p. 49), and from this Maudlin concludes that, e.g., there cannot be such a thing as strong negation. Not only is our negation not strong; in no language can negation behave that way. I cannot adequately discuss Maudlin's stance here. Suffice it to say that I agree with the observation of Gupta (2006) that Maudlin's claim does not follow from the initial assumption.

[28] Priest (1990), p. 204. [29] Ibid., pp. 204 f. [30] Ibid., pp. 205–8. [31] Ibid., p. 209.

I have two remarks to make about Priest's argument. First, note that if Priest's argument for why a dialetheist need not appeal to inexpressibility is successful, then a non-dialetheist can make use of essentially the same argument, provided she does not need for the statability of her view that the property deemed inexpressible in the object language is expressible in the metalanguage. Field's theory is arguably a non-dialetheist theory which meets this condition.[32] Second, more importantly, if we widen our perspective to look at possible languages, I think we can see that, despite Priest's claims to the contrary, there is a problem with the structure of Priest's argument. Let me explain.

Consider a hypothetical linguistic community such that the speakers of this community have implicitly made a semantic decision to use an expression to mean Boolean negation (where by their having made an 'implicit semantic decision' I simply mean that they have come to use, and come to conceive of, their negation sign as obeying the principles governing Boolean negation). Surely such a linguistic community is conceivable. And considering such a community begs no question against Priest. His argument is to the effect that Boolean negation does not exist. All that follows from this with respect to any actual or hypothetical linguistic community is that any attempt to express Boolean negation will fail.

Insofar as we are convinced by Priest's arguments that dialetheism is the correct story about our language—and suppose for argument's sake that we are—then from *our* perspective, it seems that our negation is not Boolean. Given this, an argument along Priest-style lines can be given that Boolean negation does not exist. But from the perspective of this hypothetical community, it appears that their negation is Boolean and an argument analogous to that Priest gives can be given for taking something we take to exist (perhaps truth taken as satisfying the T-schema) not to exist. What is the truth of the matter?

Here are some straightforward suggestions. (a) We are simply right and they are simply wrong (or, for that matter, they are simply right and we are simply wrong). (b) It is somehow objectively indeterminate what exists and what does not exist. (c) Both types of negation really do exist: it is just that they are expressible in different languages. (d) The underlying metaphysical assumptions ought to be questioned.

[32] Field (this volume) argues that a fully general property of determinacy—such as otherwise would cause revenge problems for his theory—in fact does not exist. The reason, briefly, is that not only does Field's object language not contain the resources to define the relevant notion of determinacy: more generally, one cannot extrapolate from the resources of the object language to make intelligible such a notion of determinacy. In the main text I focus on Priest's argument rather than Field's because Priest's argument seems more theoretically interesting. Why should Field's reasoning make us the least inclined to conclude that the relevant notion does not exist? (Scharp (this volume) discusses this problem with Field's argument at greater length. Compare too Yablo (2003), pp. 328 f.)

The suggestions are *all* very problematic. Consider first some problems with (a), which is the suggestion Priest would be forced to embrace. *First*, the situation seems clearly to be symmetric. *Second*, the consequence that whole linguistic communities might in *this* way be wrong about meaning is not easy to swallow. It is not hard to imagine cases where whole communities have mistaken *views* about what the expressions of their language mean. But a case like this would be different. Here a community would not just embrace the theory that, say, their negation sign expresses Boolean negation. Their whole use of this predicate would also point this way—they actually employ the relevant inference rules, etc.—and *yet* they could be wrong. Or that is what embracing (a) would have us accept. *Third*, setting these questions of knowledge and ignorance aside, one might think that it would be odd if meaning facts were hostage to metaphysics in the way suggested by this alternative: metaphysical facts about what propositions there are, not readily knowable to ordinary speakers, are held to determine what speakers mean. Let me elaborate on the second and third objections.

We are, to be sure, well accustomed to the idea of semantic and psychological externalism, theses to the effect that the meanings of a speaker's linguistic expressions and the contents of her mental states can be in part determined by factors external to the speaker. The arguments from Kripke, Hilary Putnam, and Tyler Burge are very familiar and taken to be persuasive.[33] Moreover, David Lewis has made popular the idea that some entities in the world are more natural, and hence more eligible to be meant and referred to, than are others.[34] This is another route to externalism. It might be thought that in light of these points, the third objection should not be very serious. But Kripkean and Putnamean appeals to the causal environment are beside the point when we are talking about expressions like 'true' and 'not'. And the appeal to the social environment in Putnam's 'elm'/'beech' case and in Burge's arthritis case is beside the point when we are talking about whole linguistic communities being wrong. The way in which the external world would matter to the meanings of speakers' expressions under hypothesis (a) is most similar to the way that Lewis takes the world to matter. A Lewisian appeal to naturalness can certainly be made also with respect to expressions that purport to refer to abstract objects. But there is a crucial difference between the kind of dependence on the external world that obtains if Lewis is right and the dependence that (a) would saddle us with. For Lewis stresses that he does not claim that *nothing* could mean *grue* or *quus*. As he notes, such a claim would on the face of it be absurd. All he claims is that if the facts about a speaker's dispositions with respect to her use of '+' do not determine whether this sign as the speaker uses it means *plus* or *quus*, then the greater naturalness of the former decides in its favor. By contrast, what we are now asked to accept is that it would be *absolutely* impossible for anyone to mean Boolean negation by a sign.

[33] See Kripke (1980), Putnam (1975), and Burge (1979).

[34] Lewis (1983) and (1984).

Someone might defend (a) from the objections pressed by attempting to argue that our intuitions about logical matters not only tell us what we do and do not mean by various logical signs, but also about what operations there are and are not in logical reality. So for instance, in so far as our intuitions are best respected by a semantics for our language given which negation is paraconsistent instead of classical, this can be taken to tell us something not only about our language but about what logical operations there are, and hence what possible languages there can be. If something like this is right, then the objection to suggestion (a) that the envisaged situation is perfectly symmetric is completely beside the point. This appeal to what members of a hypothetical linguistic community would think and intuit is as misbegotten as an objection to reliance on what our senses deliver which appeals to what hypothetical creatures whose senses would deliver contrary things would be. It is of course correct that if intuition is viewed in the way here described, the objections I have presented to suggestion (a) are beside the point. But how plausible is this view on logical intuition? (Further remarks relevant to this issue follow in the next section.)

So suggestion (a), which Priest is forced to embrace, faces serious problems. However, it is not evident that this means that Priest has been shown wrong. For the other suggestions likewise face serious problems. Turn first to suggestion (b). This is appropriately fair when it comes to the question of who is right and who is wrong. But how make sense of the postulated indeterminacy? The indeterminacy cannot be semantic, given the way semantic indeterminacy is normally understood. For in order for an expression to be semantically indeterminate as between two different meanings (the way that, for instance, a vague expression is sometimes held to be semantically indeterminate as between different precisifications) the different meanings must exist. And the problem we are currently dealing with is that the respective meanings cannot possibly coexist. So the indeterminacy in question must be ontological. But ontological indeterminacy is widely regarded with suspicion: what does it even mean for the world to be in and of itself indeterminate?[35]

As for (c), taking this route would mean abandoning Priest's universality view for a view of type (I). Moreover, given that the dialetheist's reasons for excluding Boolean negation apply equally in the case of a predicate expressing the property corresponding to Boolean negation, the concerns above stressed with respect to (I) would apply here too.

These negative reflections may make us attracted to (d): we may think that it is the associated metaphysical picture that leads us wrong in the first place. But rejecting the associated metaphysical picture—of there being these things, properties

[35] Notice too that here we are talking about a very special kind of purported ontological indeterminacy: indeterminacy in what *logical operations* there are. Postulating indeterminacy here is quite different from, say, suggesting that clouds and mountains are ontologically indeterminate.

and operations, which our predicates and logical connectives express—does not get around the problem. For what Priest is fundamentally concerned with when arguing that there is no such thing as Boolean negation is that there can be no language with an expression that functions in thus-and-such a way. This question too may be metaphysical—it concerns what languages there can be—but if it is metaphysical, it is unavoidably so. It does not presuppose a potentially objectionable ontology of entities supposedly expressed by predicates and by logical operators.

The situation is curious. Alternatives (a)–(d) appear to be exhaustive. But they all seem objectionable. This is a paradox. There is no *clean* objection here to what Priest says about Boolean negation, for although the view he is committed to, (a), seems bad, it is not clearly a worse view than the others. But Priest's position is far from unproblematic.

3.5 Possible Languages

I have discussed views (I) and (WU) on the liar paradox and expressibility, and somewhat tentatively presented some problems with both views. Let me now broaden the perspective a bit, before in the next section returning to the so-called revenge problem often raised in discussions of the liar paradox.

Consider two different questions raised by the liar reasoning:

The actual-language question: what is the correct semantic theory of 'true', 'not' and other key expressions in the liar reasoning?
The possible-language question: what possible languages are there?

The actual-language question is more often discussed. The reason for the focus on this question is clear. The liar reasoning is carried out in ordinary natural language. It presents a puzzle about how ordinary language works. Hence it is only to be expected that those theorists who seek to solve the liar paradox are concerned to get natural language right. The formal theories that are developed are meant to be accurate models of natural language. But the liar paradox also has significant implications for the possible-language question. For instance, Tarski's theorem can be regarded as a theorem about what possible languages there are. In this section I will focus on the possible-language question. For one thing, this is very much an additional question, both difficult and philosophically significant, and it is not clear that it is solved even by an otherwise perfectly adequate solution to the liar paradox. For another, consideration of the possible-language question is of consequence for the actual-language question. What if the way I take English actually to be does not correspond to the way that any language could possibly be?

What possible languages are there? Many are happy to embrace a *principle of plenitude* for abstract objects, according to which (to put things intuitively) all the pure abstracta that coherently can exist also do exist. In the case of mathematical entities, the consequences of such a principle of plenitude are *relatively* straightforward. But in the case of languages, it is harder to see what the consequences of a principle of plenitude are.[36] Take a semantic theory according to which the semantics of English is somehow fuzzy-valued (degree-theoretic). Does this semantic theory describe a possible language at all? One kind of reason for doubt is this. Even if, given a principle of plenitude, there will of course be an abstract object some of whose constitutive parts or elements—the would-be sentences—are somehow mapped onto real numbers, the question is whether this is a possible language. It is a possible language only if these real numbers can adequately be regarded as, in some sense, truth-values (or, generally, *semantic values of sentences*). This it is possible to deny, consistently with firm adherence to a principle of plenitude. A classic discussion of this is Michael Dummett's (1959), where it is argued that for general reasons having to do with the speech act of assertions, assertions can only be either correct or incorrect, and since truth and falsity corresponds to correctness and incorrectness in assertions, any appeal to 'multivalence'—there being truth-values distinct from truth and falsity—must be denied.

On what we may call a *liberal* view on what languages there are, there are languages corresponding to quite different classical and non-classical semantics. On a *restrictive* view, only languages corresponding to a very restricted class of semantics exist.

Even if a very restrictive view is correct, there are further problems with respect to what possible languages there are. For it may be—as brought up above in connection with the interlanguage paradox—that whether a given, considered by itself seemingly perfectly possible, language really exists depends on what other languages there are. For future reference, let us say that there are two possible types of restrictions on what possible languages there are: *Dummettian restrictions*, having to do with e.g. what can intelligibly be said about the truth-values of sentences (I call them 'Dummettian' since Dummett is the theorist who has done the most to put them in the foreground—but I want to include under this general label also considerations Dummett himself would refuse to endorse) and *liar-related restrictions*, having to do with what liar-type reasoning entails.

The obvious methodology for dealing with the actual-language question posed by the liar paradox is to compare proposed accounts of the paradox with intuitions we have. 'Intuitions' is used in many senses in the philosophical literature. But here it

[36] As stated in footnote 1, I here presuppose that languages are pure abstracta. If this is denied, the relevant question is instead about what languages there can be. The same considerations are still relevant.

would appear that we are dealing with *semantic* intuitions: e.g. intuitions about the truth-values of various sentences containing 'true' and other semantic predicates. The correct answer to the actual-language question posed by the liar would standardly be assumed to be: that which best respects our semantic intuitions.[37] It is certainly possible that our semantic intuitions are inconsistent. In that case, our language is arguably the language that, so to speak, is the language 'nearest' our intuitions.[38]

Attention to the possible-language question complicates this picture. First, even if our semantic intuitions were to be best respected by a theory that postulates for English a many-valued or fuzzy semantics it can be that theoretical considerations show that there are Dummettian restrictions on what languages there are, so that no language has such a semantics. Second, there are liar-related restrictions, such as illustrated by the interlanguage paradox. Can there be a language L, otherwise like a natural language, with transparent truth and satisfaction predicates?—Only if this language is not what we may call *seriously limited*: only if there is no predicate U of some possible language such that the property expressed by U cannot, on pain of paradox, be expressed in L. There can consistently be a language L of the kind described. But L cannot exist if there is another language as described. And such a language can also consistently exist.

Focus now on liar-related restrictions. Suppose that, as earlier outlined, there are two different languages L and L* both of which can consistently exist but which are incompatible. How can we know which of L and L* exists? One need not be in general skeptical of the idea of reliable substantive metaphysical intuitions to feel the force of the problem here. For even if there are reliable substantive metaphysical intuitions, problems remain. First, such intuitions do not seem to be what we actually draw on when evaluating particular proposed solutions to the liar paradox: we seem rather to rely on ordinary semantic intuitions. Second, more to the point, when it comes to choices like that between a language like English except for the possible difference that its truth predicate is transparent and a language where untruth can be expressed, is it really plausible that we have any intuitions that speak to the issue of which language it is that exists? We can perhaps have strong intuitions that speak to the issue of which

[37] I have elsewhere argued that the liar paradox shows that natural language is inconsistent, in the sense that the paradox shows that principles such that it is part of semantic competence to accept them are jointly inconsistent (See especially my (2002).) What is mentioned in the text as a possibility is something considerably weaker. There is a distinction between on the one hand semantic intuitions, in the sense merely of intuitions we actually have about semantic matters, and on the other hand what competence requires us to accept.

[38] There are problems in spelling out what 'nearness' here amounts to. But there is no way of getting around this issue. It seems rather obvious that our semantic intuitions are inconsistent—that is why the liar paradox is a paradox—and so long as that does not mean that we have failed to endow our expressions with meaning we are forced into the 'nearness' talk.

of these languages it is that *we speak*, provided both exists, but that is a different matter. Compare the case of other abstract objects; say, mathematical entities. To make this case analogous with the language case, suppose that no plenitude principle, according to which some pure mathematical entities exist if they can consistently exist, is true. Arguably, we have intuitions that promise to speak to the issue of which mathematical entities then exist. If so, then presumably what we think is that the entities quantified over in relatively natural mathematical theories like ZF and PA exist but the entities quantified over in seemingly unnatural theories do not, where perhaps Quine's NF is the prime example of an unnatural theory. Of course serious questions can be raised about the reliability of our intuitions about these matters. My point is just that in this case it can at least be reasonably urged that we *have* the relevant intuitions. By contrast, do we even have intuitions about what type of negation really exists?

If there are sufficient Dummettian restrictions, then maybe the liar-related restrictions do not present this sort of problem. For maybe there are not then two otherwise possible but incompatible languages. But in lieu of actual arguments to the effect that there are sufficient Dummettian restrictions, this is only a pious hope.[39]

3.6 Revenge Problems

Let us now return to revenge problems. Such problems are often assumed to be lethal to the purported solutions that face them. Are they?

This question is obviously closely tied to the questions that the discussion thus far has focused on, about universality and about whether certain properties simply fail to exist. There are two main ways to deal with purported revenge problems. One obvious way to deal with a revenge problem is to deny semantic self-sufficiency. This is the way of dealing with a revenge problem that corresponds to views (RI) and (I) on the

[39] Let me also stress an independent reason to be concerned with the interlanguage paradoxes. If indeed our semantic intuitions are inconsistent, as suggested above, and the correct semantic theory of (the relevant fragment of) English is just the theory that does the best job of capturing these inconsistent intuitions, then the following possibility should easily suggest itself: it is simply semantically indeterminate what is the semantic value of 'true' (and of other key expressions in the liar reasoning and its variants). There are different semantic theories assigning different semantic values to 'true', such that nothing about whatever determines the semantic value of 'true' determines between these theories. The point is naturally put in terms of possible languages. Nothing determines, among these possible languages, which one it is that we speak. There is a range of languages such that, for all that is determined about what is our actual language, any one of them could be our natural language. Call these possible languages 'candidate-languages'. The relevance of the interlanguage paradoxes is that they show that we cannot take for granted that these candidate-languages all exist; perhaps at most one does, in which case the semantic values of the relevant expressions can be determinate after all.

expressive power of natural language. One problem with such a strategy is that it can seem that the relevant property clearly is expressible already in English. (Consider a view according to which *truth* is not expressible in English, or according to which it is not expressible in English that a sentence is *unsettled or false*.) Call this type of problem *incredulity*. A second problem is that of whether, for instance because of the reasoning of the interlanguage paradox discussed earlier, it just is not coherent to maintain that some property is expressible only in another language: for the relevant property may anyway be indirectly expressible, and that is sufficient for problems to arise. Call this second problem *the interlanguage problem*. The interlanguage problem arises for some, but by no means all, views of this general kind. Given that general arguments for semantic self-sufficiency like those of McGee and Priest are flawed, I do not see what other problem besides these two could threaten to arise generally for this type of way of dealing with the revenge problem. In fact, I suspect that even those discussions of revenge problems that on the face of it focus on self-sufficiency are motivated by what I called the incredulity problem: the problem is not ascent into a richer metalanguage per se but rather that the property said to be expressible only in this other language is one that we seem already to be able to express.

Those who want to avoid having to kick up properties into the metalanguage because of the liar paradox—that is, those who do not accept views (RI) or (I)—seek to show that no revenge problems arise for their theories. Normally this is done by an attempt to demonstrate that the expressive resources employed in stating the solution can be allowed into the object language with no untoward consequences: that our language can be taken to be self-sufficient. Suppose this is successfully done for some given theory of the liar. This does not necessarily avoid either the incredulity problem or the interlanguage problem. Take first the incredulity problem. Even if none of the expressions employed in a particular account of the liar has to be conceived of as belonging to a richer metalanguage, the theory can still have the consequence that some property that intuitively seems expressible in English in fact is not expressible. So the incredulity problem can still arise.[40] What is more, even if the theory's account of the truth and satisfaction predicates is consistent and even if no untoward consequences follow from allowing the expressive resources employed in stating the solution into the object language, the resources for an interlanguage liar paradox may still be available, so long as (i) there is some language with a predicate expressing a property which on pain of contradiction cannot be expressed by any predicate in the object language, and (ii) such a property is thereby indirectly expressible already in the object language.

Let me close by making a few brief remarks on the incredulity problem. One might think that the incredulity problem is anyway the really hard one: that all otherwise

[40] See again the discussion of Field's theory of truth in Scharp (this volume).

acceptable theories of the liar paradox end up deeming inexpressible in a given natural language properties that clearly are expressible in that language. There are two ways to respond to this suspicion. One is of course to attempt to devise a theory which does not deem any intuitively expressible properties inexpressible. A second, more original and theoretically more involved suggestion involves being more careful about expressibility. Consider the kind of view on the liar paradox I have myself elsewhere defended—an *inconsistency view*, I will here call it.[41] According to this view, the liar paradox arises because meaning-constitutive principles for expressions of our language are inconsistent, where a meaning-constitutive principle is something it is part of semantic competence with the relevant expressions to be disposed to accept.[42] I will not here attempt to defend this sort of view. But what I wish to stress is that this sort of view provides the materials for a promising response to the incredulity problem.

On an inconsistency view one can say that the meaning-constitutive principles for the logical connectives are that they satisfy the inference rules characteristic of classical logic and a meaning-constitutive principle for the truth predicate is that it satisfy the T-schema. Tarski's theorem then shows that the semantic values of the relevant expressions cannot possibly make true all the relevant meaning-constitutive principles. The semantic values of the relevant expressions are then what come closest to satisfying the relevant meaning-constitutive principles, possibly given other constraints. Suppose—I am only concerned to give a 'model' here—that the truth predicate fails to satisfy the T-schema. It is then fairly natural to say, absurd though it may sound, that *truth fails to be expressible in English*. There is no predicate of English that expresses what the truth predicate *aims to express*; no predicate that satisfies the meaning-constitutive principles for the truth predicate. And even if saying that this means that truth fails to be expressible in English is unnecessarily paradoxical, that does not affect the main point. The inconsistency view provides the materials for a way of getting around incredulity problems. We can respect the idea that truth obviously is expressible in English by noting that there is a predicate of English which aims to express it (or, to cash this out, which is governed by the right meaning-constitutive principles). We can respect the result that truth does not seem to be expressible in English by noting that (on the assumptions mentioned) there is no predicate of English whose semantic value satisfies the meaning-constitutive principles for 'true'.

Let an expression *misfire* iff its semantic value fails to satisfy the associated meaning-constitutive principles. The point is then that on an inconsistency view some expressions will misfire. Specifically, one can think that there can be properties such

[41] See Eklund (2002).

[42] There are many important matters concerning formulation that I here slide over. For instance, it is far from clear that this talk of *dispositions to accept* is what best describes what semantic competence involves.

that any expression of a given language that aim to express this property will misfire. Such a property is in the most straightforward sense not expressible in the language in question. But while the property is inexpressible there may be predicates that aim to express it.

Notice that if this is how the incredulity problem is dealt with, it is far from clear why a view of type (I) should be preferable to a view of type (RI). For the reason that a type (I) view would be preferable is that it is more palatable that a seemingly 'technical' property should be expressible only in a richer language than that an 'ordinary' property which ordinary speakers seem clearly able to express should be expressible only in a richer language. But given the route just outlined, a property deemed inexpressible is one that at the same time is claimed that we aim to express. And it is far clearer that we *aim* to express an 'ordinary' property such as that of being true than that we actually *succeed* in this aim.[43]

References

Barwise, Jon, and Etchemendy, John (1987). *The Liar: An Essay on Truth and Circularity*, Oxford University Press, Oxford

Beall, JC (ed.) (2003). *Liars and Heaps: New Essays on Paradox*, Clarendon Press, Oxford

Beall, JC, and Armour-Garb, Bradley (eds.) (2005). *Deflationism and Paradox*, Oxford University Press, Oxford

Belnap, Nuel (1982). 'Gupta's rule of revision theory of truth', *Journal of Philosophical Logic* 11: 103–16

Burge, Tyler (1979). 'Individualism and the mental'. In Peter French, Theodore Uehling, and Howard Wettstein (eds.), *Midwest Studies in Philosophy* IV, University of Minnesota Press, Minneapolis, pp. 73–121

Eklund, Matti (2002). 'Inconsistent languages', *Philosophy and Phenomenological Research* 64: 251–75

Field, Hartry (2003a). 'A revenge-immune solution to the semantic paradoxes', *Journal of Philosophical Logic* 31: 1–27

____ (2003b). 'The semantic paradoxes and the paradoxes of vagueness', in Beall (2003)

____ (2005a). 'Is the liar sentence both true and false?', in Beall and Armour-Garb (2005)

[43] Note that if the view sketched in the last few paragraphs is accepted, and view of type (RI) is acceptable, then many of the statements in this chapter will have to be taken with a 'grain of salt' (compare Frege (1892), p. 192): for some property that I aim to express will not in fact be expressible. But in contrast with other theorists who ask not to be begrudged a pinch of salt, I have a theoretical explanation of what is going on when my expressions do not in fact express what they are meant to express.

An earlier version of this chapter was presented as a paper at a conference on paradoxes at the University of Latvia, November 2005. Thanks to the conference participants for good discussion. I also wish to thank Dan Korman, Agustín Rayo, and Kevin Scharp for helpful comments on earlier versions of this chapter.

Field, Hartry (2005b). 'Variations on a theme by Yablo'. In Beall and Armour-Garb (2005)
_____ this volume, 'Solving the paradoxes, escaping revenge'
Fitch, Frederic (1946). 'Self-reference in philosophy', *Mind* 55: 64–73
_____ (1964). 'Universal metalanguages for philosophy', *Review of Metaphysics* 17: 396–402
Frege, Gottlob (1892). 'On concept and object'. In Michael Beaney (ed.), *The Frege Reader*, Blackwell, Oxford (1997), pp. 181–93
Gupta, Anil (1982). 'The liar paradox', *Journal of Philosophical Logic* 11: 1–60. Reprinted in Martin (1984)
_____ (1997). 'Definition and revision: a response to McGee and Martin'. In Villanueva (1997), pp. 419–43
_____ (2006). '*Truth and Paradox: Solving the Riddles*, by Tim Maudlin', *Mind* 115: 163–5
Gupta, Anil, and Belnap, Nuel (1993). *The Revision Theory of Truth*, MIT Press, Cambridge, Mass.
Herzberger, Hans (1970). 'Paradoxes of grounding in semantics', *Journal of Philosophy* 67: 145–67
_____ (1982). 'Notes on naive semantics', *Journal of Philosophical Logic* 11: 61–102. Reprinted in Martin (1984)
_____ (1982a). 'Naive semantics and the liar paradox', *Journal of Philosophy* 79: 479–97
Hofweber, Thomas (2006). 'Inexpressible properties and propositions'. In Dean Zimmerman (ed.), *Oxford Handbook of Metaphysics*, vol. 2
Kripke, Saul (1975). 'Outline of a theory of truth', *Journal of Philosophy* 72: 690–716. Reprinted in Martin (1984)
_____ (1980). *Naming and Necessity*, Harvard University Press, Cambridge, Massa.
Lewis, David (1983). 'New work for a theory of universals', *Australasian Journal of Philosophy* 61: 343–77
_____ (1984). 'Putnam's paradox', *Australasian Journal of Philosophy* 62: 221–36
Martin, Robert L (1976). 'Are natural languages universal?', *Synthèse* 32: 271–91
_____ (1984). *Recent Essays on Truth and the Liar Paradox*, Clarendon Press, Oxford
_____ (1984a). 'Introduction'. In Martin (1984), pp. 1–8
Maudlin, Tim (2004). *Truth and Paradox: Solving the Riddles*, Clarendon Press, Oxford
McGee, Vann (1991). *Truth, Vagueness and Paradox*, Hackett Publishing Company, Indianapolis
_____ (1994). 'Afterword: truth and paradox'. In Robert M. Harnish (ed.), *Basic Topics in the Philosophy of Language*, Prentice Hall, Englewood Cliffs, NJ, pp. 615–33
_____ (1997). 'Revision', in Villanueva (1997), pp. 387–406
Parsons, Terence (1990). 'True contradictions', *Canadian Journal of Philosophy* 20: 335–53
Priest, Graham (1984). 'Semantic closure', *Studia Logica* 43: 117–29
_____ (1987). *In Contradiction*, Kluwer Academic Publishers, Dordrecht
_____ (1990). 'Boolean negation and all that', *Journal of Philosophical Logic* 19: 201–15
_____ (1995). 'Beyond gaps and gluts: reply to Parsons', *Canadian Journal of Philosophy* 25: 57–66
_____ (2005). *Doubt Truth to Be a Liar*, Oxford University Press, Oxford
Putnam, Hilary (1975). 'The meaning of "meaning" '. In Keith Gunderson (ed.), *Language, Mind, and Knowledge*, University of Minnesota Press, Minneapolis, pp. 131–93
Scharp, Kevin: this volume, 'Aletheic vengeance'
_____ manuscript, 'Truth and internalizability'

Simmons, Keith (1993). *Universality and the Liar*, Cambridge University Press, Cambridge

Soames, Scott (1999). *Understanding Truth*, Oxford University Press, Oxford

Tarski, Alfred (1935/83). 'The concept of truth in formalized languages'. In John Corcoran (ed.), *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, 2nd edn, Indianapolis: Hackett Publishing Company. English translation by J. H. Woodger of 'Der Wahrheitsbegriff in Formalisierten Sprachen', *Studia Philosophica* 1 (1935)

Villanueva, Enrique (ed.) (1997). *Truth*, Ridgeview, Atascadero, Calif.

Yablo, Stephen (2003). 'New grounds for naive truth theory'. In Beall (2003), pp. 312–30

4

Solving the Paradoxes, Escaping Revenge

Hartry Field

It is 'the received wisdom' that any intuitively natural and consistent resolution of a class of semantic paradoxes immediately leads to other paradoxes just as bad as the first. This is often called the 'revenge problem'. Some proponents of the received wisdom draw the conclusion that there is no hope of *any* natural treatment that puts all the paradoxes to rest: we must either live with the existence of paradoxes that we are unable to treat, or adopt artificial and *ad hoc* means to avoid them. Others ('dialetheists') argue that we can put the paradoxes to rest, but only by licensing the acceptance of some contradictions (presumably in a paraconsistent logic that prevents the contradictions from spreading everywhere).[1]

I think the received wisdom is incorrect. In my effort to rebut it, I will focus on a certain *type* of solution to the paradoxes. This type of solution has the advantage of keeping the full Tarski truth schema

$$\textbf{(T)} \quad \mathit{True}(\langle A \rangle) \leftrightarrow A$$

[1] This latter view is only reasonable if 'revenge' is less of a worry for inconsistent solutions to the paradoxes than for consistent ones. I think myself that advocates of inconsistent solutions face a prima-facie revenge problem, and doubt that they can escape it without employing the devices I suggest in this chapter on behalf of certain consistent solutions. But that is a matter for another occasion.

(and more generally, a full satisfaction schema). This has a price, namely that we must restrict both the law of excluded middle and the law connecting $A \rightarrow B$ with $\neg A \vee B$, but we can carve the restrictions narrowly enough so that ordinary reasoning (e.g. in mathematics and physics) is unaffected.[2] I'll call solutions of this type *G-solutions*. (If you want to think of the 'G' as standing for 'good' I won't stop you.) The literature contains several demonstratively consistent solutions of this sort; for purposes of this chapter there is no need to choose between them. (I will give an informal introduction to this type of solution in sections 1, 3, and 4, and a formal account in section 5.) It will turn out that any such solution generates certain never-ending hierarchies of sentences that may seem 'increasingly paradoxical' (roughly speaking, it is harder to find a theory that satisfactorily treats later members of the hierarchy than to find one that satisfactorily treats earlier members); but the G-solution gives a consistent treatment of each member of each such hierarchy. The existence of these hierarchies prevents certain kinds of revenge problems from arising: certain attempts to state revenge problems simply involve going up a level in a hierarchy all levels of which have been given a non-paradoxical treatment.

Still, there are certain 'vindictive strategies' (strategies for trying to 'get revenge' against solutions to the paradoxes) that G-solutions may seem to be subject to. I'll argue that the most popular such strategy is based on a misunderstanding of the significance of model-theoretic semantics. But there is a far more interesting strategy for which this is not so. As mentioned, a G-solution generates certain never-ending hierarchies of apparently paradoxical sentences which however are each successfully treated by the account. But shouldn't it be possible to 'break out of the hierarchies' to get paradoxes that are not resolved by the account? Or to put it another way: If we *can't* 'break out of the hierarchies' *within the language that our solution to the paradoxes treats*, isn't that simply due to an expressive limitation in that language? That, I think, is the most difficult revenge worry for G-solutions to deal with.

Some of the worries about 'breaking out of the hierarchies' turn out to be intimately connected to König's paradox of the least undefinable ordinal, a paradox in the same ballpark as those of Berry and Richard. The G-solutions provide a consistent treatment of these definability paradoxes. The treatment of König's paradox will be an important element in my argument that we are unable to 'break outside the hierarchies', but that this does not reflect an expressive limitation of the language.

[2] Also, the equivalence between $A \rightarrow B$ and $\neg A \vee B$ will hold on the assumption of excluded middle for A and for B.

Part One Introductory Discussion

1 The paradoxes and excluded middle

Imagine that we speak a first-order language *L*: it has the usual connectives and quantifiers, and it contains no ambiguous terms and no indexicals. It is to be a very rich language, powerful enough to express all our mathematics, including the richest set theory we currently know how to develop. It should be able to talk about its own expressions and their syntax; though we needn't actually make this a separate requirement, since as Gödel showed we can use arithmetical surrogates. If we like we can also assume that *L* can express all our current claims about the physical world too, though this will not really matter to the problem to be discussed. Finally, *L* should contain terms like 'true' and 'true of'. For present purposes we needn't worry about how such terms apply to sentences and formulas in languages other than our own, so we may as well assume that they have been restricted to apply only to the sentences and formulas of *L*.

These assumptions about our language *L* are enough to generate paradoxes (or rather, apparent paradoxes). Some of them, like the Liar paradox, arise from the fact that by any of a number of well-known routes we can construct self-referential sentences: sentences that attribute to themselves any property you like. For instance, the Liar paradox arises from any sentence that directly or indirectly asserts its own untruth; let Q be some such sentence, and $\langle Q \rangle$ its standard name.[3] Since Q asserts its own untruth, it certainly seems that

$$Q \leftrightarrow \neg True(\langle Q \rangle)$$

had better be part of our overall theory. In addition, it seems that our theory of truth ought to include every 'Tarski biconditional' , i.e. every instance of the schema (T) mentioned earlier; hence in particular,

$$True(\langle Q \rangle) \leftrightarrow Q.$$

[3] There is a familiar distinction between *contingent* and *non-contingent* Liar sentences. If the sentence 'Nothing written on the first blackboard manufactured in 2005 is true' is written on a blackboard that (perhaps unbeknownst to the writer) was the first to be manufactured in 2005, it is a contingent Liar: given the contingent facts about blackboard-manufacture it in effect asserts its own untruth. Non-contingent Liar sentences assert their own untruth independent of such empirical facts. Some of the formulations below are only strictly correct for non-contingent Liars, but could easily be generalized to apply to contingent Liars too. (For instance, in the next sentence of the text, replace 'be part of our overall theory' by 'follow from our overall theory together with the empirical facts'.) The distinction between the two kinds of Liar sentences will make no important difference.

But then if the conditional, and the biconditional defined from it in the obvious way, are at all reasonable, we can infer

(*) $True(\langle Q \rangle) \leftrightarrow \neg True(\langle Q \rangle)$.

And being of form $B \leftrightarrow \neg B$, this leads to contradiction in classical logic.

There are similar paradoxes that don't require the construction of self-referential sentences. For instance, just as our theory of truth ought to include the instances of Schema (T), so our theory of satisfaction ought to include the instances of the following schema:

(S) For all x, x satisfies $\langle A(v) \rangle$ if and only if $A(x)$

(where to say that x satisfies $\langle A(v) \rangle$ is the same as saying that $\langle A(v) \rangle$ is true of x).[4] In the special case where $A(v)$ is the formula 'v does not satisfy v', this yields that for all x,

$\langle v$ does not satisfy $v \rangle$ satisfies x if and only if x does not satisfy x,

and hence

(**) $\langle v$ does not satisfy $v \rangle$ satisfies itself if and only if it does not satisfy itself. [5]

Again, (**) is of form $B \leftrightarrow \neg B$ and hence leads to contradiction in classical logic.

The 'G-solutions' that I'll be considering accept these derivations of (*) and (**). But unlike 'dialetheic' views (e.g. [16]), they do not accept contradictions (sentences of form $C \wedge \neg C$). So they must reject all arguments that would take us (for arbitrary B) from $B \leftrightarrow \neg B$ to a sentence of form $C \wedge \neg C$.

I think that the most revealing way of trying to argue from $B \leftrightarrow \neg B$ to a contradiction is as follows:

(i) Assume both $B \leftrightarrow \neg B$ and B. Then by modus ponens, $\neg B$; so $B \wedge \neg B$.

(ii) Assume both $B \leftrightarrow \neg B$ and $\neg B$. Then by modus ponens, B; so $B \wedge \neg B$.

[4] (S) should really be called (S_1): it is the schema for the satisfaction predicate 'satisfies$_1$' that applies to formulas with exactly one free variable, and there is an analogous schema (S_n) for each satisfaction predicate 'satisfies$_n$' that applies to formulas with exactly n free variables. But (S_1) can be taken as basic: in any language rich enough to code finite sequences we can artificially define the higher satisfaction predicates in terms of 1-place satisfaction, in a way that guarantees the schemas for the former if we have (S_1): e.g. to say that 'v_1 is larger than v_2' is satisfied by o_1 and o_2 in that order is in effect to say that 'u is an ordered pair whose first member is larger than its second' is satisfied by $\langle o_1, o_2 \rangle$.

We can similarly reduce truth to satisfaction: to say that 'Snow is white' is true is in effect to say that 'Snow is white and $u = u$' is satisfied by everything (or equivalently, by something). (T) then falls out of (S_1), so (S_1) can be taken as the sole basic schema. But it's more natural to talk in terms of truth than satisfaction, so I'll keep on talking about (T).

[5] A natural abbreviation for 'v satisfies itself' would be 'v is onanistic'. But for some reason 'v is homological' has caught on instead, with 'heterological' for 'non-onanistic'.

(iii) Since $B \wedge \neg B$ follows both from the assumptions $B \leftrightarrow \neg B$ and B and from the assumptions $B \leftrightarrow \neg B$ and $\neg B$, then it follows from the assumptions $B \leftrightarrow \neg B$ and $B \vee \neg B$. (Reasoning by cases.)

(iv) But $B \vee \neg B$ is a logical truth, so $B \wedge \neg B$ follows from $B \leftrightarrow \neg B$ alone.

I now further stipulate that G-solutions accept both modus ponens and reasoning by cases (*aka* disjunction elimination). So they take the reasoning to be valid through step (iii).

What G-solutions question is the use of the law of excluded middle in step (iv). Unlike intuitionists, though, G-theorists take excluded middle to be perfectly acceptable within standard mathematics, physics, and so forth; it is only certain reasoning using truth and related concepts that are affected.[6] There is a verbal issue here about exactly how this point should be put. One way to put it is to say that excluded middle is literally *valid* in some contexts like mathematics, but invalid outside that domain. But it might be thought that the 'topic neutrality' of logic implies that if excluded middle can't be accepted everywhere then it can't be taken as literally *valid* anywhere. Even so, this doesn't undermine the claim that it is *effectively* valid[7] in contexts like mathematics: if one accepts all instances of the schema $A \vee \neg A$ that don't contain 'true', then even if one doesn't claim that they are *logical* truths one can reason from them just as a classical logician reasons in mathematics and physics. So it really makes no difference in which of the two ways we talk.

Another argument from $B \leftrightarrow \neg B$ to a contradiction runs as follows: after step (i) as above, we conclude that $B \leftrightarrow \neg B$ entails $\neg B$ by a *reductio* rule (that if X and B together entail $\neg B$, then X alone entails $\neg B$); that result and (ii) then give the contradiction. But the most obvious argument for that reductio rule is based on the law of excluded middle (together with reasoning by cases). The argument is that if X and B together entail $\neg B$, then since X and $\neg B$ certainly entail $\neg B$, it follows that X and $B \vee \neg B$ entails $\neg B$; and since $B \vee \neg B$ is a logical truth, this means that X entails $\neg B$. So I will

[6] Actually advocates of G-solutions *might* want to further restrict excluded middle, e.g. by disallowing its application to certain sentences containing vague concepts; and indeed it is not out of the question to regard certain mathematical concepts such as 'ordinal' as having a kind of 'indefinite extensibility' that is akin to vagueness. Still, for purposes of this chapter I assume that excluded middle applies unrestrictedly within standard mathematics.

Another plausible restriction of excluded middle is to sentences containing normative concepts like 'appropriate' or 'reasonable'; this is relevant to certain 'doxastic paradoxes' involving, for instance, sentences asserting that it is not appropriate to believe them. But such paradoxes are outside the scope of this chapter.

[7] 'Effectively valid' means 'in effect valid': it has nothing to do with effective procedures. Similarly I'll use 'effectively classical' to mean 'in effect classical', i.e. excluded middle holds even if not as a logical law.

assume that in giving up (or restricting) excluded middle we give up (or restrict) this *reductio* rule as well.

Admittedly, this *reductio* rule is valid in intuitionist logic even in absence of excluded middle, so I can't say that we are *compelled* to give up the reductio rule if we give up excluded middle. But intuitionist logic does not evade the paradoxes, so we had best not follow its lead.[8] My point is that there is a natural response to the Liar paradox which sees this kind of *reductio* reasoning as depending on the law of excluded middle and both as needing restriction; and that is the response that G-solutions adopt.

2 Trying to preserve classical logic

Weakening classical logic to deal with the paradoxes is obviously not something to be done lightly, and there are questions about how to understand the proposal, some of which I will address in the next section. But first I'd like to briefly survey the options for handling the paradoxes within classical logic. One reason for doing this is to make evident the appeal of the non-classical approach, and another is to facilitate a later discussion of the 'hierarchies of paradoxical sentences' that arise within G-solutions.

In classical logic, the reasoning of the Liar paradox can easily be turned into a proof of the following disjunction:

Either
(i) $\langle Q \rangle$ is true, but $\neg Q$
or
(ii) $\langle Q \rangle$ is not true, but Q.

At this point, classical theorists have three options. (Of course, there is also the possibility of remaining agnostic between the options, but that is of no particular interest.)

The first option is to choose disjunct (i). This would seem quite unattractive: doesn't calling $\langle Q \rangle$ true while saying 'nonetheless, $\neg Q$' deprive the notion of truth of significance?

The second option is to choose disjunct (ii). This seems on its face almost equally unattractive: if one holds that $\langle Q \rangle$ is not true, what is one doing holding Q?

[8] Intuitionists tend to motivate the *reductio* rule by way of the law $\neg(A \wedge \neg A)$ (sometimes misleadingly called the 'law of non-contradiction'). But to anyone who accepts the deMorgan law $\neg(A \wedge B) \models \neg A \vee \neg B$, this version of the 'law of non-contradiction' simply amounts to $\neg A \vee \neg\neg A$, a slightly restricted version of excluded middle that few who reject excluded middle would accept. (That's why dialetheists who accept excluded middle accept $\neg(A \wedge \neg A)$, making clear that it does not adequately capture the principle that we should reject contradictions.) The intuitionist argument for *reductio* thus turns on their rejection of the deMorgan law.

The third option is to accept the disjunction of (i) and (ii) *while ruling out as absurd the acceptance of either disjunct*. (It is because the acceptance of either disjunct is viewed as absurd that this is really a third option, distinct from agnosticism between the first two options). This third option takes the acceptance of either (i) or (ii) to be absurd, on the ground that commitment to A *requires* commitment to A being true and conversely; but it nonetheless allows commitment to $A \vee \neg A$. Now, many people think that if one accepts a disjunction of two options each of which would be absurd to accept, one has already accepted an absurdity. Indeed, that principle appears to be built into classical logic: it is the principle of reasoning by cases (or disjunction elimination), to which attention was called above. This third option is based on rejecting that principle, except in restricted form.[9] So it is probably best thought of as only a *semi*-classical option: it does accept all the validities of classical logic, but disallows natural applications of disjunction elimination and some of the other standard meta-rules.

These three options seem to be the only possibilities for keeping the validities of classical logic without accepting contradictions.[10] Admittedly, one could insist with Tarski that the predicate 'true' should be given a hidden subscript, or that its extension

[9] The restricted form is that if Γ together with A entail C *by classical rules*, and Γ together with B entail C *by classical rules*, then Γ together with $A \vee B$ entail C. The third option can accept that, but cannot accept the generalization to 'entailment' by the truth rules (that commitment to A requires commitment to $\mathit{True}(\langle A \rangle)$ and conversely). And these truth rules must have a quasi-logical status on the third option, since it was only by holding acceptance of (i) and of (ii) to be *absurd* that the view differentiated itself from agnosticism between the first two options.

[10] I know of no one who has seriously proposed taking the first option. Classical and semi-classical logicians who do technical work on the paradoxes mostly tend to prefer the third option: see [10] and [14]; also [9], where five of the nine types of theories discussed fall under option three. The option of choice among non-specialists seems to be option two, but some specialists prefer it as well, e.g. [2] and [13]. (If the description of the latter as a classical theory seems surprising, see [8].)

What about Kripke's seminal [11]? That's more complicated since Kripke offers a model-theoretic semantics with no instructions on how to read the theory off the semantics. But if we interpret him as suggesting that though the extension of 'True' is a fixed point, the logic is classical, then his theory also falls under option two.

An alternative and I think more attractive interpretation of Kripke is to take the set of acceptable sentences to coincide with the extension of 'True': they are both the contents of the same fixed point. But if the fixed points are based on a Kleene semantics, this gives a non-classical logic, and so is not germane to the discussion in this section. (This way of interpreting Kripke has been advocated in [21]—not altogether consistently, in my view, since Soames talks in terms of truth-value gaps, which seems to presuppose the classical logic interpretation. [19] clearly distinguishes the two ways of getting a theory from a Kripkean fixed point, in the distinction between the theories there called KF and KFS.)

On the non-classical reading of Kripke, his solution is similar in sprit to the G-solutions under discussion in this chapter; however, the nonclassical logic one obtains from this way of reading Kripke is unsatisfactorily weak, since Kleene semantics has no serious conditional. G-solutions do much better in this regard.

vary with context. Still, given classical logic (even in the weak sense that includes only the validities and not the meta-rules), the above three options are the only consistent ones *when the subscript and context are held fixed*.[11]

A problem with all of the classical and semi-classical solutions is that they prevent the notion of truth from fulfilling its generally accepted role. The standard story about why we need a notion of truth ([18], [12]) is that we need it to make certain kinds of generalizations. For instance, the only way to generalize over

$$\begin{array}{l}(\text{Snow is white}) \rightarrow \neg\neg(\text{Snow is white})\\ (\text{Grass is green}) \rightarrow \neg\neg(\text{Grass is green})],\end{array}$$

is to first restate them in terms of truth and then generalize using ordinary quantifiers:

For every sentence, if it is true, so is its double negation.

But this says what we want it to say only if we assume the intersubstitutivity of $True(\langle A\rangle)$ with A: that is, the principle

Intersubstitutivity: If Y results from X by replacing some occurrences of A with $True(\langle A\rangle)$, then X and Y entail each other. [This needs to be restricted to cases where the substitution is into contexts that aren't quotational, intentional, etc.; but I'll take the language L to contain no such contexts.]

This principle entails the truth schema in classical logic, indeed in any logic in which $A \leftrightarrow A$ is a logical truth. So the three classical and semi-classical options all reject the intersubstitutivity principle. Thus they fail to satisfy the purpose of the notion.

For instance, relative to the assumption that what Jones said was exactly $A_1, \ldots, A_n$, we want

If everything Jones said is true then___

to be equivalent to

If A_1 and . . . and A_n then___.

This requires that the $True(A_i)$ be intersubstitutable with the A_i inside the conditional, but that won't in general be so on *any* of the classical and semi-classical theories. The semi-classical theory does *better* than the fully classical ones: it allows for intersubstitutivity of $True(\langle A\rangle)$ with A in more contexts. Indeed the fully classical ones don't even allow substitutivity for unembedded occurrences: $True(\langle A\rangle)$ and A can't be mutually entailing in a classical theory that includes disjunction elimination (as we'll see in the next section). But though the semi-classical theories do *better*, that

[11] I'm putting aside solutions to the Liar paradox based on unmotivated syntactic restrictions that prevent the formation of self-referential sentences. Very strong syntactic restrictions are required for this, and the solutions are of little interest since they do not generalize to the heterologicality paradox.

isn't good enough. An advantage of G-solutions is that not only do they accept the Tarski schema, they accept the full Intersubstitutivity Principle.

I conclude this section with some further remarks on the second classical option; in particular, on a version of the second classical option that invokes a hierarchy of truth predicates. This will play a role later in the chapter, where I will compare it to a hierarchy of strengthenings of a single truth predicate that arises in G-solutions.

A common theme among proponents of the second option is that the schema (T) holds for all sentences that 'express propositions', where to 'express a proposition' is to be either true or false, i.e. to either be true or have a true negation. (On this view, expressing a proposition is much stronger than being meaningful: it would be hard to argue that the 'contingent Liar sentences' of note 3 aren't meaningful, but proponents of this option take them not to express propositions.) So instead of (T) we have

$$\text{(RT)} \quad [\mathit{True}(\langle A\rangle) \vee \mathit{True}(\langle \neg A\rangle)] \rightarrow [\mathit{True}(\langle A\rangle) \leftrightarrow A].$$

It is easily seen that this is equivalent to the left-to-right half of (T), i.e. to

$$\text{(LR)} \quad \mathit{True}(\langle A\rangle) \rightarrow A.^{12}$$

Obviously, then, a decent theory of truth containing (RT)/(LR) needs to contain vastly more. (It's compatible with (RT)/(LR) that nothing is true; or that only sentences starting with the letter 'B' are true; etc.) The crucial question for such a theory is, how are we to fill it out without leading to paradox?

It turns out that however we try to fill it out, we are led to the conclusion that *basic principles of the truth theory itself* fail to be true. (They also fail to be false, so that they come out as 'not expressing propositions' .) Of course any non-contingent Liar sentence is itself an assertion of the theory that the theory asserts not to be true, but it presumably doesn't count as one of the basic principles of the theory. But what does count as a basic principle of the theory is (RT), or its equivalent (LR). And Montague [15] showed that with very minimal extra assumptions, one can derive from (LR) a conclusion of form

$$\neg \mathit{True}[\langle \mathit{True}(\langle M\rangle) \rightarrow M\rangle],$$

i.e. that some specific instance of (LR) isn't true. Most people regard it as a serious defect of a theory that it declares central parts of itself untrue; and saying that these parts 'don't express propositions' doesn't appear to help much.

This is the point at which the idea of a hierarchy of truth predicates may suggest itself. The idea is that we don't have a general truth predicate, but only a hierarchy of predicates 'true$_\alpha$', where the subscripts are notations for ordinal numbers (in a

[12] In proving a given instance of schema (RT) from schema (LR) one uses two instances of the latter, one for A and one for $\neg A$.

suitably large initial segment of the ordinals that has no last member). The Montague theorem then shows that the principles of the theory of truth$_\alpha$ aren't true$_\alpha$, but the idea is to try to ameliorate this by saying that they're true$_{\alpha+1}$. Call such a view a *stratified truth theory*.

Besides their artificiality, stratified truth theories seriously limit what we can express, in a way that undermines the point of the notion of truth. For instance, suppose we disagree with someone's overall theory of something, but haven't decided *which part* is wrong. The usual way of expressing our disagreement is to say: not all of the claims of his theory are true. Without a general truth predicate, what are we to do? The only obvious idea is to pick some large α, and say 'Not all of the claims of his theory are true$_\alpha$' . But this is likely to fail its purpose since we needn't know how large an α we need. (Indeed, there would be strong pressure on each of us to use very high subscripts α even in fairly ordinary circumstances, but however high we make it there is a significant risk of it not being high enough to serve our purposes. This was the lesson of the famous discussion of Nixon and Dean in [11]. Nixon and Dean wanted to say that nothing the other said about Watergate was true, and to include those assertions of the other in the scope of their own assertions; but to succeed, each needed to employ a strictly higher subscript than the other.)

Actually, the situation is even worse than this. For suppose that we want to express disagreement with a stratified truth theorist's overall 'theory of truth' (i.e. the theory he expresses with all of his 'true$_\alpha$' predicates), but that we haven't decided which part of that theory is wrong. Here the problem isn't just with knowing how high an α to pick; rather, *no* α that we pick could serve its purpose. The reason is that it's *already part of the stratified theory* that some of its claims aren't true$_\alpha$, namely, the principles about truth$_\alpha$; that's why the theorist introduced the notion of truth$_{\alpha+1}$. So *whatever* α we pick, we won't succeed in expressing our disagreement.

The problems just mentioned are really just an important special case of a problem that I've argued to infect all classical and semi-classical theories: they can't give truth its proper role as a device of generalization. Except possibly for dialetheic theories, which I will not consider here, restricting excluded middle seems to be the only way to avoid crippling limitations on our notion of truth.

3 More on rejecting excluded middle

It is important to note that in classical logic you don't need anything like the full strength of the truth schema (T) (or the satisfaction schema (S)) to derive contradictions: indeed, if you allow reasoning by cases as well as the classical validities, all that is required is the two assumptions

(T-Elim) A follows from *True*($\langle A \rangle$)

and

(T-Introd) $True(\langle A\rangle)$ follows from A

(or the analogous Elimination and Introduction rules for satisfaction). For using these instead of (T), we can easily recast the derivation (i)-(iv) in Section 1 (with $True(\langle Q\rangle)$ as the B) as follows:

(i*) By (T-Elim), $True(\langle Q\rangle)$ implies Q,[13] which is equivalent to $\neg True(\langle Q\rangle)$; hence $True(\langle Q\rangle)$ implies the contradiction $True(\langle Q\rangle) \wedge \neg True(\langle Q\rangle)$;

(ii*) $\neg True(\langle Q\rangle)$ is equivalent to Q, which by (T-Introd) implies $True(\langle Q\rangle)$; hence $\neg True(\langle Q\rangle)$ also implies the contradiction $True(\langle Q\rangle) \wedge \neg True(\langle Q\rangle)$.

(iii*) Since $True(\langle Q\rangle) \wedge \neg True(\langle Q\rangle)$ follows both from the assumption $True(\langle Q\rangle)$ and from the assumption $\neg True(\langle Q\rangle)$, then it follows from the assumption $True(\langle Q\rangle) \vee \neg True(\langle Q\rangle)$. (Reasoning by cases.)

(iv*) But $True(\langle Q\rangle) \vee \neg True(\langle Q\rangle)$ is a logical truth, so we have a derivation of the contradiction $True(\langle Q\rangle) \wedge \neg True(\langle Q\rangle)$.

(If we strengthened (T-Elim) to the assumption of the conditional $True(\langle A\rangle) \rightarrow A$, we could give a derivation that doesn't involve reasoning by cases.)

In fact, we don't even need the full strength of (T-Introd); we can make do with the weaker assumption

(T-Incoherence) A and $\neg True(\langle A\rangle)$ are jointly inconsistent.

Inconsistency proof: $True(\langle Q\rangle)$ implies Q by (T-Elim), and $\neg True(\langle Q\rangle)$ implies Q since it is equivalent to Q, so we derive Q using reasoning by cases plus excluded middle. Using the other half of the equivalence between Q and $\neg True(\langle Q\rangle)$, we get $Q \wedge \neg True(\langle Q\rangle)$, which is inconsistent by (T-Incoherence).

The fact that the paradox arises from weaker assumptions than (T) is important for two reasons. First and most obviously, it means that if we insist on keeping full classical logic we must do more than restrict (T), we must restrict the weaker assumptions as well. But the second reason it's important concerns not classical solutions, but G-solutions: it gives rise to an important moral for what G-solutions have to be like.

For even though G-solutions take truth to obey the Tarski schema (T), we'll see that they recognize other 'truth-like' predicates (e.g. 'determinately true') that don't obey the analog of (T) but do obey the analogs of (T-Elim) and (T-Introd) (or at the very least, (T-Elim) and (T-Incoherence)). For each truth-like predicate, there is a Liar-like sentence that asserts that it does not instantiate this predicate.

[13] That is, Q follows from $True(\langle Q\rangle)$. (One reader took my 'A implies B' to mean 'if A then B', and on this basis accused me here of illicitly extending (T-Elim) to hypothetical contexts; but that is not what I mean by 'implies'.)

Reasoning as in (i*) and (ii*) is thus validated, and since G-solutions accept reasoning by cases without restriction, paradox can only be avoided by rejecting the application of excluded middle to these Liar-like sentences formed from truth-like predicates. In short, G-solutions are committed to the view that *there can be no truth-like predicate for which excluded middle can be assumed*. (Since excluded middle is to hold within ordinary mathematics and physics, this means that no truth-like predicates can be constructed within their vocabulary.) As we'll see, the conviction that there *must* be truth-like predicates obeying excluded middle is one primary source of revenge worries.

I close this section by trying to make clear what is involved in restricting the application of excluded middle to certain sentences, e.g. the Liar sentence, when one accepts the intersubstitutivity of $True(\langle Q \rangle)$ with Q. In particular, what is the appropriate attitude to take to the claim $True(\langle Q \rangle)$? According to the sort of solution to the paradoxes I've sketched, one must reject the claim that $True(\langle Q \rangle)$ and also reject the claim that $\neg True(\langle Q \rangle)$, since these claims each imply a contradiction relative to any theory of truth that implies the Tarski biconditionals. (One can take rejection as a primitive state of mind, involving at the very least a refusal to accept; a slightly more informative account of rejection can be found in [4] (Section 3).) We must likewise reject the corresponding instance of excluded middle

$$Z: \quad True(\langle Q \rangle) \vee \neg True(\langle Q \rangle),$$

for it too leads to contradiction. And because we reject Z, our refusal to either accept $True(\langle Q \rangle)$ or accept $\neg True(\langle Q \rangle)$ doesn't seem appropriately described as 'agnosticism' about the truth of Q. We would be agnostic about $True(\langle Q \rangle)$ if we believed Z but were undecided which disjunct to believe; but when we reject Z the very factuality of the claim that $True(\langle Q \rangle)$ is being put into question, so our not believing $True(\langle Q \rangle)$ while also not believing $\neg True(\langle Q \rangle)$ isn't happily described as 'agnosticism' .

It should be immediately noted that a solution of this sort does *not* postulate a 'truth-value gap' in Q: it does not say that Q is neither true nor false, i.e. that neither Q nor its negation is true. It also does not say that Q is neither true nor *not true*. Saying that Q is 'gappy' or 'non-bivalent' in either of these senses would trivially entail that Q is not true, which (by the Tarski biconditionals and modus ponens) leads to contradiction. *Since the claim that Q is 'gappy' (non-bivalent) leads to contradiction, we must reject it.*

That isn't to say that we should believe that Q is bivalent (or that it is not 'gappy'; these are the same, assuming the equivalence of $\neg\neg A$ to A, as I henceforth shall). The claim that Q is bivalent or non-gappy amounts to

$$Z^*: \quad True(\langle Q \rangle) \vee True(\langle \neg Q \rangle).$$

This in turn amounts to Z (non-truth and falsity turn out to coincide as applied to sentences in the language), and as we've seen, Z must be rejected.

If it seems odd that both the claim that Q is gappy and the claim that it is not gappy lead to contradiction, it shouldn't: from the fact that $Gappy(\langle Q \rangle)$ and $\neg Gappy(\langle Q \rangle)$ each lead to contradiction, all we can conclude is that

$$Z^{@}: \quad Gappy(\langle Q \rangle) \vee \neg Gappy(\langle Q \rangle)$$

leads to contradiction; so the proper conclusion is that this instance of excluded middle must also be rejected.[14] In other words, *the claim that Q is 'gappy' has the same status as Q itself has*. In particular, just as it is misleading to declare ourselves 'agnostic' about the Liar sentence, it is also misleading to declare ourselves 'agnostic' about the claim that the Liar sentence is 'gappy' or the claim that it is bivalent: for we don't recognize that there is a fact to be agnostic about.

I think it would be a serious problem if there were no way to assert the 'defective' status of Q within the language. As we'll see, there is a way; but it can't be done by saying that Q suffers a truth-value gap.

4 The Berry–Richard–König paradox

I think that all of the semantic paradoxes turn on excluded middle, though some of them (especially some of the ones involving conditionals) do so in an indirect and unobvious fashion. I will make this precise in Section 5. There I will introduce a language that contains a 'quasi-classical conditional' which obeys many of the classical laws for conditionals even in the absence of excluded middle; moreover it reduces to the material conditional when excluded middle is assumed for antecedent and consequent. I will then state a result (proved elsewhere) according to which every semantic 'paradox' *that can be formulated in this language* has a solution that is compatible with the Tarski biconditionals. The solution may depend on the failure of some of the classical laws for the conditional, but that failure will always be traceable to a breakdown in excluded middle for the antecedent or consequent of one of the conditionals in question. We thus diagnose these apparent paradoxes as only apparent, they depend on illicit applications of excluded middle.

Of course, the fact that those apparent paradoxes *that can be formulated in the language* turn out not be genuinely paradoxical does not settle the revenge issue: settling that issue requires considering the possibility of expanding the language to get new paradoxes. I will have a lot to say toward undermining the idea of revenge in later sections.

[14] Indeed whenever one rejects a given instance $A \vee \neg A$ of excluded middle, one should also reject the instance $(A \vee \neg A) \vee \neg(A \vee \neg A)$, for they are equivalent by very uncontroversial reasoning; hence one should reject $Bivalent(\langle A \rangle) \vee \neg Bivalent(\langle A \rangle)$. [Reason for the equivalence: $\neg(A \vee \neg A)$ implies $\neg A$, so $(A \vee \neg A) \vee \neg(A \vee \neg A)$ implies $(A \vee \neg A) \vee \neg A$, which implies $A \vee \neg A$. The other direction is trivial.]

First though I will consider how the paradoxes of definability fare on this sort of view. There are a number of slightly different paradoxes of definability, the most famous being Berry's and Richard's, but they all have the same underlying idea. Because of its relevance later in the chapter, I will focus attention on a variant of the Berry and Richard paradoxes due to König.

Recall that *L* is a first order language adequate to expressing its own syntax, and that contains a satisfaction predicate. From that predicate we can define '*L*-definable':

> *z* is *L*-definable if and only if there is at least one formula of *L* (with exactly one free variable) that is satisfied by *z* and by nothing else.

Now, *L* is assumed to be built from a finite or countably infinite vocabulary, so it contains only countably many formulas; from which it follows that only countably many things are *L*-definable. But there are uncountably many ordinal numbers; indeed, uncountably many *countable* ordinal numbers. So there are (countable) ordinal numbers that are not *L* -definable. So there is a smallest ordinal number that is not *L*-definable, and it must be unique. But then '*v* is an ordinal number that is not *L*-definable but for which all smaller ordinals are *L*-definable' is uniquely satisfied by this ordinal, so it is *L*-definable after all, which is a contradiction. That is the paradoxical line of argument.

Any solution to the paradoxes of satisfaction will implicitly contain a solution to this definability paradox. On *classical logic* solutions, if the language *L* contains the predicate 'satisfies' then certain instances of schema (S) from Section 1 are refutable; and in particular, if we define '*L*-definable' from 'satisfies' as above, there will be counterinstances to even the more restricted schema

(S_{defn}) For all *x*, *x* satisfies '*v* is *L*-definable' if and only if *x* is *L*-definable.

This gives one possible diagnosis of the error in the argument: that it lies in the inference from σ being the uniquely smallest *L*-undefinable ordinal to '*v* is the smallest *L*-undefinable ordinal' being uniquely satisfied by σ.

But on the approach that I've sketched, we are committed to maintaining all instances of (S), and in particular all instances of (S_{defn}). Where then does the reasoning of the paradox go wrong?

Where the reasoning goes wrong, I think, is that it makes an implicit application of excluded middle to a formula involving '*L*-definable'. Excluded middle can be assumed for certain restricted definability predicates. For instance, let L_0 be obtained from *L* by deleting 'satisfies' and terms defined from it (such as 'definable') or closely related to it; then excluded middle holds for formulas that contain 'L_0-definable' (as long as they don't contain problematic terms in addition). There is no even *prima facie* problem about the least ordinal that is not L_0-definable, since the description of it just given is in a part of *L* that goes beyond L_0. Similarly for expansions of L_0 in which the application of 'satisfies' is somehow restricted in a way that guarantees excluded

middle (e.g. a language L_1 in which 'x satisfies y' can occur only in the context 'x satisfies y and y is an L_0-formula', or a language L_2 in which 'x satisfies y' can occur only in the context 'x satisfies y and y is an L_1-formula'). Given a well-defined hierarchy of such expansions, each of which includes all the vocabulary of the previous, one gets within L a hierarchy of restricted definability predicates, each more inclusive than the previous. But for definability in the full language L, the fact that excluded middle must be rejected for 'satisfies' suggests that it will almost certainly have to be rejected for the predicate 'L-definable' defined from it; and the paradox shows that indeed it does.

The implicit application of excluded middle to a formula involving 'L-definable' occurred in the step from

(1) There are ordinal numbers that are not L-definable

to

(2) There is a smallest ordinal number that is not L-definable.

To see that the inference from (1) to (2) depends on excluded middle, consider any specific ordinal β, and suppose that every ordinal less than β is L-definable. Given this supposition, (2) says in effect

(3) Either β is not L-definable, or there is an ordinal $\alpha > \beta$ such that α is not L-definable and all its predecessors *are* L-definable.

But this entails

(4) β is not L-definable or β is L-definable;

and so if we reject (4) we must reject (2).[15] But there is no call to reject (1): there are certainly ordinals that are not L-definable, for uncountable ones can't be L-definable (and there may well be sufficiently large countable ones which are definitely not L-definable too). So the inference from (1) to (2) relies on excluded middle.

This resolution of the paradox may seem to have a high cost. For the inference from 'There are ordinals α such that $F(\alpha)$' to 'There is a smallest ordinal α such that $F(\alpha)$' is absolutely fundamental to ordinary set-theoretic reasoning; doesn't what I'm saying count as a huge and crippling restriction on ordinary set theory? Not at all: ordinary set theory allows sets to be defined only by 'effectively classical' properties, that is, properties F for which the generalized law of excluded middle $\forall x[F(x) \vee \neg F(x)]$ holds. I'm not suggesting any restriction whatever on the ordinary laws of set theory; what I am saying, and what is independently quite obvious, is that one has to be very careful if one wants to *extend* set theory by allowing properties (or formulas) that aren't known to be effectively classical into its axiom schemas.

[15] Which isn't to say that we should accept the negation of (2): that would require (an existential quantification of) a negation of excluded middle, which would lead to contradiction.

This point is worth elaboration. Standard set theory (ZFC) contains two axiom *schemas* (the schemas of Separation and Replacement). On a *strict* interpretation of the theory, the allowable instances of the schemas are just those instances in the language of set theory; however, the 'impure' set theory that most of us accept and employ is more extensive than this, it allows instances of the schemas in which physical vocabulary occurs (e.g. we take the separation schema to allow us to pass from the existence of a set of all non-sets to the existence of a set of all neutrinos). But when the law of excluded middle is not assumed to hold unrestrictedly, there is a question of just how far the extension should go. I think a suitable extension of the schema of separation to be the rule

(Extended Separation) $(\forall x \in z)(Ax \vee \neg Ax) \vdash \exists y \forall x(x \in y \leftrightarrow x \in z \wedge A(x))$

(allowing free parameters in the formula $A(x)$), *where any vocabulary at all, including 'true', can appear in* $A(x)$. Requiring excluded middle as an assumption of separation seems reasonable. Otherwise, we would license sets for which membership in the set depends on whether the Liar sentence is true; given extensionality, this would lead at the very least to indeterminate identity claims between sets, and it isn't at all clear that paradox could be avoided even allowing that. But Extended Separation as formulated above avoids such oddities, while allowing such sets as the set of true sentences in the 'true'-free set-theoretic language; it seems to me as much of an extension of separation to the language containing 'true' as we ought to want.

It is easy to see that if the formula $F(x)$ is allowed to contain non-classical vocabulary, then Extended Separation (together with the fact that every non-empty set has a member of least rank) justifies reasoning from 'There is at least one ordinal α such that $F(\alpha)$ *and such that for all ordinals* $\beta < \alpha$, $F(\beta) \vee \neg F(\beta)$' to 'There is a smallest ordinal α such that $F(\alpha)$' . And in applications of set theory in which the formula $F(x)$ is in standard mathematical or physical vocabulary, we don't need to bother stating the italicized clause since it is always satisfied. But once we allow $F(x)$ to contain notions like 'true' or 'satisfies' or notions explained in terms of them such as '*L*-definable', that clause is required: forgetting it involves an illicit assumption of excluded middle, and it is on that that the Berry-Richard-König paradox rests.[16]

[16] What I've said here about the 'least ordinal principle' is true of the least number principle in arithmetic: it too requires an excluded middle premise. Also, it's worth remarking that even positive forms of induction (ordinary or transfinite) can in general only be assumed in rule form, e.g.

$$\forall \alpha[(\forall \beta < \alpha)(F\beta) \rightarrow F\alpha] \models \forall \alpha F\alpha.$$

An excluded middle premise is required for the conditional form

$$\forall \alpha[(\forall \beta < \alpha)(F\beta) \rightarrow F\alpha] \rightarrow \forall \alpha F\alpha.$$

But the rule form suffices for typical applications.

Part Two Model Theory and Revenge

5 Conditionals and G-logics

It's now time to give slightly more detail about the sort of logic I have in mind for dealing with the paradoxes—a *G-logic*, I'll call it. My plan is not to specify any one such logic, but to specify a class of logics any of which would deal with the paradoxes along the lines I have sketched. The logics differ only in the details of the treatment of the conditional. (I warn the reader that this section is somewhat technical; a cursory skim will probably suffice for understanding most of what follows.)

The simplest way to specify the class of G-logics is to specify a type of model-theoretic semantics for members of this class—a *G-semantics*. For any specific G-logic £, the corresponding G-semantics will give a definition of £-*valid inference* within classical set theory; since classical set theory is accepted both by the advocates of £ and their classical opponents, the definition of £-validity will be intelligible to all.

I need to make a small generalization of the usual framework for model-theoretic definitions of validity: I need to allow the size of the valuation space on which the model is based to depend on the size of the domain of the model. More fully, for any cardinal number c, we fix a value space V_c with a specific subset D_c (the ' designated values' of V_c); the various V_c may all be the same, but needn't be. We then stipulate that a *c-model M* consists of a non-empty domain *U of cardinality no greater than c* for the quantifiers of the language to range over, together with an assignment of an object in U to each name of the language, an operation on U to each function symbol in the language, and a 'V_c-valued extension' to each predicate in the language; where a V_c-valued extension (for an n-place predicate) is a function that assigns members of V_c to n-tuples of members of U. This apparatus (in conjunction with certain operations on V_c) will be used to assign a value $|A|_M$ in V_c to each sentence A of the language ('the semantic value of A in M'). We then define an inference among sentences of the language to be *c-valid* (in the given logic £) if in *every* c-model in which the premises take on designated values of V_c, the conclusion does too. And we define it to be *valid* if it is c-valid for every cardinal number c. (The definition extends in a natural way to inferences among formulas with free variables.) As remarked above, this notion of validity is definable in classical set theory, which is something that advocates of £ and advocates of classical logic both accept. So it is understood, and understood *in the same way*, by all concerned parties.

The idea of defining validity in a model-theoretic semantics *formulated in classical set theory* may seem to make it inevitable that some sort of paradox will arise for the account. For we've seen above that a G-solution can't allow for 'truth-like' predicates to which excluded middle applies. But isn't 'has a designated value' a 'truth-like' predicate? And doesn't defining it in classical set theory guarantee that excluded

middle must hold for it? It may seem, then, that if we are going to pursue a G-solution to the paradoxes we must explain the logic in some other way than by a model theory given in classical set-theoretic terms—say, by a model theory given in 'the non-classical part of' the G-language (in which case one would have to bootstrap one's way into an understanding of the logic plus model theory package). I think that this supposed problem for explaining the logic within classical set theory is illusory: it rests upon a misunderstanding of the nature of model theory.[17] But I defer the argument for this to Sections 8 and 9.

Turning now to the specifics of the model theory for G-logics, what should we take the value spaces V_c to be, and how should we assign values in them to sentences? If it weren't for the conditional, we could use a simple 3-valued semantics (whatever the cardinality of the model): the values might be called 0, $\frac{1}{2}$, and 1, and we would assign one of these values to each sentence; or more generally to each formula relative to an assignment s of objects in the domain of M to its free variables. (The values are of course all relative to M as well as to s, and indeed, the set of possible s depends on the domain of M; but for convenience I'll omit the reference to M, and often the reference to s as well, in what follows.) We'd take the value of a conjunction (relative to s) to be the minimum of the values (relative to s) of the conjuncts, the value of a disjunction the maximum, and the value of $\neg A$ to be 1 minus the value of A. The quantifiers are analogous to conjunction and disjunction. (More precisely, the value of $\forall x A$ relative to s is the minimum of the value of A relative to all the various expansions of s obtained by assigning objects in the domain of M to x.) We'd take 1 as the sole designated value: that is, we'd take the valid inferences to be those that in all valuations preserve the value 1. Then no sentence and its negation can both be designated; and instances of excluded middle needn't be designated, since they can have value $\frac{1}{2}$. This is called the *strong Kleene semantics* for the conditional-free language, and the logic for the conditional-free language that it generates is called *Kleene logic*. It's the semantics Kripke mostly used in [11].

This simple approach won't work if the language is to contain a conditional validating the schemas (T) and (S), for it is not hard to see that there is no 3-valued connective behaving anything like a conditional for which the associated Tarski biconditionals all can have value 1.[18] But the approach can be generalized: we can let the spaces V_c have many more than three values (indeed, for each cardinal c we can take V_c to have more than c values), and we can take each V_c to be only

[17] I don't deny that there might be a place for a non-classical model theory to supplement the classical one, or even to supplant it for those already converted to the G-logic.

[18] The 3-valued conditional that 'comes closest' to adequacy is the Lukasiewicz 3-valued conditional (where $|A \rightarrow B|$ is 1 if $|A| \leq |B|$, 0 if $|A| = 1$ and $|B| = 0$, and $\frac{1}{2}$ otherwise). But even here, if C is a Curry-like sentence that asserts that if it is true then so is the Liar sentence Q, then the Tarski biconditionals for Q and C can't both get value 1.

partially ordered instead of linearly ordered. I will consider only the case where each V_c has a least element 0 and a greatest element 1. I will assume that there is an 'up-down symmetry' operation * on V_c: an operation that reverses order and which when applied twice to any element leads back to that element. This operation will correspond to negation. I will also require that V_c be 'c-complete' : that is, each set of members of V_c that has cardinality no greater than c must have a greatest lower bound and a least upper bound. (In all cases of interest the models are infinite, so c is infinite; so 'c-completeness' implies '2-completeness' , i.e. every pair of elements has a greatest lower bound and least upper bound. If we wanted to consider the case where $c < 2$, we'd have to add a 2-completeness requirement.) Using c-completeness (and 2-completeness), we can take the value of $A \wedge B$ to be the greatest lower bound of the values of A and of B, and the value of $\forall x A$ to be the greatest lower bound of the values of A relative to all the possible assignments of objects to x; similarly for $\vee$ and $\exists$, using least upper bounds.

We also want to guarantee that as in the 3-valued semantics, $\vee$-Elimination and $\exists$-elimination hold. I will stick to the simplest case, in which 1 is the sole designated value.[19] Then the most natural way to guarantee the validity of $\vee$-Elimination is to stipulate that 1 is 'join-irreducible' , i.e. that the least upper bound of two elements of V_c isn't 1 unless one of those elements is 1. This guarantees $\vee$-Elimination, in the restricted form that if the inferences from A to C and from B to C are both valid then so is that from $A \vee B$ to C. And we can get from this to the more general form (that if the inferences from Γ and A to C and from Γ and B to C are both valid then so is that from Γ and $A \vee B$ to C) if we assume the distributive law; for this and other reasons, I will assume distributivity. Similar remarks apply to $\exists$-elimination: for the restricted form without side formulas Γ, we assume that 1 is 'c-join-irreducible' , i.e. that if S is any subset of V_c whose cardinality is no greater than c, then 1 is not the least upper bound of S unless 1 is a member of S;[20] and we get the unrestricted form that allows for side formulas if we make a weak infinite distributivity assumption.[21]

Because of the features mentioned so far, the spaces V_c can be regarded as a 'fine-graining' of the 3-valued semantics: in any such space, the values other than 0 and 1 simply partition the value $\frac{1}{2}$ given in the 3-valued semantics. As a result,

[19] A more general approach would take the set of designated values to be any prime filter not containing both v and v^* for any v, where * is the operator corresponding to negation.

[20] In typical cases $V_c - \{1\}$ will have no maximum member. (Indeed, we'll see some reason in Section 10 to impose a condition (∂c) on a satisfactory G-semantics that would entail this.) In such cases, 1 must be the least upper bound of $V_c - \{1\}$, so the condition of c-join irreducibility implies that V_c must have cardinality greater than c and hence greater than the cardinality of any model that has it for a value space. This is why I allow the value space to depend on an upper bound of the cardinalities of the models that employ it.

[21] Namely $a \sqcap (\sqcup_\alpha \{b_\alpha\}) = \sqcup_\alpha \{a \sqcap b_\alpha\}$, when $\{b_\alpha\}$ has cardinality no greater than c.

when considering inferences among conditional-free sentences it makes no difference whether you use the 3-valued semantics or one of the V_c. So the logic governing conditional-free sentences is just Kleene-logic.

The point of the fine-graining is to handle the conditional. What we need to do is add an operator $\Rightarrow$ on the spaces V_c of fine-grained values to correspond to the $\rightarrow$. The operator should obey reasonable laws, of which the foremost is

(I) $A \rightarrow B$ should have value 1 when and only when the value of A is less than or equal to that of B.[22]

Defining $A \leftrightarrow B$ as $(A \rightarrow B) \wedge (B \rightarrow A)$, (I) implies that $A \leftrightarrow B$ has value 1 if and only if A and B have the same value. Given that all the operators are evaluated value-functionally, this then implies

($\mathbf{I}_{Cor}$) When $A \leftrightarrow B$ has value 1 and X_B results from X_A by substituting B for one or more occurrences of A then X_B should have the same value as X_A.

I'll say that two formulas A and B are *of equal strength* if $A \leftrightarrow B$ is valid, and that A *is at least as strong as* B if $A \rightarrow B$ is valid. Because of (I), the claim that A is at least as strong as B amounts to the claim that for every model M and every c at least as big as the domain of M, $|A|_M \leq |B|_M$ in V_c. So when A is at least as strong as B, the inference from A to B is valid; the converse claim fails.

Further reasonable laws for the conditional include the following:

(II) Strengthening of the consequent should strengthen the conditional (not necessarily strictly), and strengthening of the antecedent should weaken it;

(III) If A has value 1 and B has value 0 then $A \rightarrow B$ should have value 0;

and probably

(IV) $A \rightarrow B$ should have the same value as $\neg B \rightarrow \neg A$.

(I) (together with the assumptions we've made about negation) already implied a weak form of (IV), viz. that $A \rightarrow B$ should have the value 1 if and only if $\neg B \rightarrow \neg A$ does; the extension (IV) seems highly natural, but will play only a tangential role in what follows.

Note that (I) and (III) together imply that when the values of A and B are restricted to the set $\{0, 1\}$ (i.e. when $A \vee \neg A$ and $B \vee \neg B$ take on value 1), then the conditional $A \rightarrow B$ is to be evaluated just like the material conditional $\neg A \vee B$, which given the

[22] Note that the validity of the inference from A to B does not guarantee that when A has value less than 1, its value is less than or equal to that of B; so given (I), we can't in general conclude from the validity of the inference from Γ, A to B that Γ implies $A \rightarrow B$. We can conclude this in the special case where Γ implies $A \vee \neg A$.

restriction is bound to behave classically. So *any failure of a classical law for the conditional is ultimately due to a failure of excluded middle in an antecedent or consequent.*

We might expand this list of reasonable laws in various ways, but for the sake both of generality and simplicity let's leave it at that;[23] then a *DMC-semantics* ('deMorgan semantics with conditional') is any semantics based on partially ordered sets V_c with operators that satisfy the laws mentioned. There are many examples of DMC-semantics in the literature: the most famous examples (besides the 2-valued semantics for classical logic) are the semantics for the various Lukasiewicz multivalued logics. (In the Lukasiewicz logics, the same value space—e.g. the interval of real numbers—is used for every cardinal c; and the partial order $\leq$ is in fact a linear order.)

Unfortunately, most DMC-semantics will not suffice for our needs. Indeed, we've seen that no semantics consistent with the truth schema (T) can permit a 'truth-like' predicate that obeys excluded middle; but in most versions of DMC-semantics, including the Lukasiewicz versions, such predicates can be defined using the conditional and 'True'.

Our overall goal is a DMC-semantics that is consistent with the truth schema (T), or more generally the satisfaction schema (S). Given (I_{Cor}), this requires that there be models in which for each sentence A, $True(\langle A \rangle)$ has the same value as A; and for each formula A with one free variable, $Satisfies(x, \langle A \rangle)$ always has the same value as the result A_x of replacing all free occurrences of the free variable in A by 'x'.[24] That's what *consistency* with (T) and (S) require. Actually what we want is more than mere consistency, we want a kind of 'conservativeness' result involving 'consistency with any standard starting model' . Basically what this requires is that any standard 'starting model' (classical model M_0 for the fragment L_0 not containing 'True' and 'Satisfies') can be converted to a model of the DMC semantics that meets these conditions on 'True' and 'Satisfies' and whose part not involving 'True' and 'Satisfies' 'looks just like' M_0.[25] (Whenever in the rest of the chapter I talk of consistently adding the truth schema

[23] For later reference I mention a strengthening of (III):

(III_s) If A has greater value than B then $A \rightarrow B$ should have value 0.

If the values were linearly ordered, this with (I) would yield excluded middle for conditional claims, which would inevitably breed paradox; but I know of no system adequate to the paradoxes whose values are linearly ordered, and the published G-solutions do all satisfy (III_s).

[24] With suitable change of bound variables in A, if x occurs in A so as to create a conflict with the substitution.

[25] By a *standard* model I mean one whose syntactic part is standard (i.e. for each e_2 in the domain there are only finitely many e_1 in the domain for which $\langle e_1, e_2 \rangle$ satisfies 'v_1 and v_2 are expressions and v_1 is part of v_2'); equivalently, whose arithmetic is standard. The 'looks just like M_0' condition means that the two models have the same domain and the same assignments to individual constants and function symbols, and that M assigns to any n-place predicate of L_0 a function that maps any n-tuple into 1 if it is in the M_0-extension of the predicate and into 0 otherwise.

to a semantics, what I really mean is this.) Let a *G-semantics* be any DMC-semantics that meets this conservativeness requirement.

There are in the literature several different G-semantics: several solutions to the paradoxes of the general sort I've sketched that give all the biconditionals of form (T) and (S) value 1.[26] Given (I_{Cor}), this means that *they also validate the intersubstitutivity of True*($\langle A \rangle$) *with A, even within the scope of other operators such as the conditional*: if X results from Y by substituting *True*($\langle A \rangle$) for one or more occurrences of A, then X and Y get the same value; so $X \rightarrow Y$ and its converse get value 1. (More generally, they validate the intersubstitutivity of *Satisfies*(x, $\langle A \rangle$) with A_x even within the scope of other operators. In what follows I will frequently state my claims just for truth, leaving the generalization to satisfaction tacit.) Logics that can be based on such semantics (*G-logics*) are the ones that will be of interest in what follows.

For purposes of this chapter it will be convenient to use the term 'valid' in a very broad sense, one which counts every arithmetic truth, a large amount of set theory, and the basic principles of truth and satisfaction as 'valid'. To be explicit, let a *quasi-correct model of the 'true'-free fragment* L_0 of L be a model M_0 of L_0 such that for some inaccessible cardinal κ, if *Ur* is the subset of the domain that does not satisfy 'Set', then the set-theoretic portion of M_0 consists of the set of all (not necessarily pure) sets of rank less than κ built from urelements in *Ur* (together with the usual $\in$ relation on this domain). Let a *standard set-theoretic model for L* (in a given G-semantics) be a model of L with valuation space V_c for which the model for the 'true'-free fragment is a quasi-correct model of cardinality no greater than c, and in which all instances of the schemas (T) and (S) get value 1 (when syntax is developed in ZFC in a standard way). Then I will take an inference to be *valid* in the G-logic if it preserves the designated value 1 in any standard set-theoretic model of the G-logic. For future reference, I note that the following three set-theoretic principles come out valid in this sense even when extended to the full language containing 'True' and 'Satisfies': (i) the Extended Separation Schema of Section 4; (ii) the rule form of transfinite induction, mentioned in note 16, and (iii) the 'choice principle' $(\forall x \in X)(\exists y)F(x, y) \models (\exists f)[dom(f) = X \wedge (\forall x \in X)F(x, fx)]$ (from which a rule form of the Replacement Schema follows).[27]

[26] One is almost explicit in [3] and fully explicit in [5]; more mathematical details about it can be found in [22]. Another can be found in [6]; this one was inspired by [24], which is in the general spirit of a G-semantics but fails to satisfy intersubstitutivity of logical equivalents within conditionals. One can also modify Lukasiewicz continuum-valued semantics to get a G-solution, using the basic ideas from [3]. I suspect that there are many other possibilities as yet undiscovered. A much earlier paper offering something close to a G-semantics is [1].

[27] (i) and (ii) are completely evident given the quasi-correctness of the underlying model. (iii): if the premise is to have value 1 in the model, then using the axiom of choice it must be that for some function g with domain $\{o \mid ||x \in X||_o = 1\}$, $||F(x, y)||_{o,g(o)} = 1$ for all $o \in dom(g)$. By the quasi-correctness of the

6 Semantic values and truth

How do the semantic values in a G-semantics relate to the notion of truth? Putting aside a complication to be discussed in the next section, we can say that the following holds for 'reasonable' models:[28]

- Sentences with value 1 are true;
- Sentences with value 0 are false (i.e. have true negations).

These claims are natural given the semantics. If a sentence has value 1 then anyone who knows that it has this status can assert it; and since the claim that A is true is equivalent to A itself, he or she can then assert that it is true. Similarly, anyone who knows the status of a sentence with value 0 can assert that it is false: for its negation must have value 1, so we can assert the truth of $\neg A$ and hence the falsity of A. This is intended only as an informal argument; that is the best that can be expected for connecting notions of these two different kinds.

How about sentences that (we know to) have intermediate values? It is sometimes assumed that they are neither true nor false, but that does not fit with what I've already stipulated, in particular with the intersubstitutivity of $True(\langle A\rangle)$ with A. We've seen this already with Liar sentences: they must have an intermediate value, but we can't assert that they aren't true since that would lead to contradiction, so we certainly can't assert that they aren't either true or false. But the point holds more generally. Since falsehood is equivalent to truth of the negation, to say of a sentence A that it is neither true nor false would be to say

$$\neg True(\langle A\rangle) \wedge \neg True(\langle \neg A\rangle).$$

But by the intersubstitutivity of $True(\langle A\rangle)$ with A, that would be equivalent to saying

$$\neg A \wedge \neg\neg A,$$

i.e. to accepting a contradiction; which is illegitimate in this logic since contradictions can never get the only designated value, viz. 1. So *even for sentences for which we can easily show that they can't have value 0 or 1*, we must reject the claim that they are neither true nor false. (Of course if we know them to have intermediate semantic value, we also won't assert that they are either true or false; nor will we assert that they are either true, false, or neither true nor false.)

underlying model (plus Replacement), this g must be in the model, from which it follows that the conclusion must get value 1.

[28] We'll see there that *even in classical semantics*, the conditions may fail for some sentences whose quantifiers range over sets of arbitrarily high rank. You can take the discussion in this section to apply only to sentences with suitably bounded quantifiers.

Is there anything we *can* say about the truth and falsity of sentences with intermediate semantic values? Yes. Some examples:

- When the value of A is less than or equal to that of B, then if A is true B is true and if B is false A is false.
- If A and B are both true, so is $A \wedge B$, and if $A \vee B$ is true then at least one of A and B is true.[29]

The claims in the second bullet each result directly from three applications of the equivalence between $True(\langle C \rangle)$ and C (together with the fact that '$\wedge$' and '$\vee$' are transcriptions of 'and' and 'or'). And those in the first bullet can be reduced to the two that were informally argued for two paragraphs back. For if the value of A is less than or equal to that of B, then the value of $A \rightarrow B$ is 1; so by the claim of two paragraphs back, $A \rightarrow B$ is true, and hence (by the intersubstitutivity properties of truth and the fact that '$\rightarrow$' means 'if . . . then') if A is true then so is B. Similarly for the falsehood claims, using the law (IV) for the conditional. (This is the only place in the chapter I will rely on (IV).)

In sum, there is a lot we can say about the connection between the semantic values and truth and falsity; but we can't say of any claim with intermediate value either that it is true or that it isn't. Again, it wouldn't be appropriate to say that we're ignorant about whether these sentences are true. For we are ignorant about whether A is true when either A is true or A is not true but we don't know which; but in this case the assumption that either A is true or A is not true cannot be made.

7 Revenge problems: introductory remarks

Since one of the requirements of a G-logic was that it be consistent with the instances of (T) and (S) all having value 1 (or rather, the stronger 'conservativeness' requirement sketched near the end of section 5), there is no danger that there are any genuine paradoxes statable in the language: any apparent paradox statable in the language has a solution that is consistent with the instances of these schemas. Thus a vast array of apparent paradoxes (e.g. a Curry-like paradox involving a sentence C that asserts that if it's true then so is the Liar sentence) are automatically solved. But it might be argued that the language is expressively weak, in that certain notions that we can't easily do without are inexpressible in it; and that including those notions within the language would inevitably breed new paradoxes. This is the general idea behind revenge problems.

[29] It is tempting to summarize these two bulleted laws by saying that truth is a 'fuzzy prime filter' on the space of values—'fuzzy' because it is indeterminate whether any given sentence with intermediate truth value is in it.

One notion to worry about is determinacy. We've seen that we can't without contradiction declare that the Liar sentence isn't true; but we nonetheless reject the Liar sentence (since it leads to contradiction), and it seems that there is some important sense in which we believe that it is not *determinately* true. But if we recognize such a sense of determinacy, then a full solution to the paradoxes must consider sentences that can include this notion as well as 'True' and '$\rightarrow$' (and all the other notions in the original language).

I agree with this, and will discuss how a G-logic can accommodate it in Section 10. Before that, though, I want to consider one particular form of the worry that I do *not* agree with. This worry involves a particular kind of notion of determinacy, one that is thought to be somehow read off the model-theoretic semantics. In Section 9 I will give two 'simple revenge arguments' based on this form of the worry, and argue that they are mistaken. But first I will prepare the way for my reply, by trying to remove what I think are common misconceptions about model-theoretic semantics.

Not all revenge arguments are based on misconceptions about model theory: I will discuss what I see as a much more interesting revenge argument, not based on such a misconception, in Sections 13-20. I think there is some tendency in discussion of these matters for the 'simple' and 'sophisticated' arguments to become intertwined, so it is important to deal with the simple ones before dealing with the sophisticated ones.

8 Model-theoretic semantics

In Section 5 I sketched a model-theoretic semantics for the language L, in classical set theory. What is the value of giving such a semantics? The obvious answer, and the one I gave, is that such a semantics enables us to give a set-theoretic definition of a notion of logical validity (equivalently, of logical implication) for the language. When, as here, the language of set theory is part of L and is assumed to effectively obey classical logic (by virtue of $A \vee \neg A$ always being assumed, when A is in the language of set theory), we are using what is in effect a classical part of L to define validity for the full L.

That it is possible to adequately develop the theory of validity (implication) for L within a classical portion of L rests on a presupposition. It presupposes that excluded middle holds of logical notions like implication: that is, it presupposes that if Γ is a set of sentences and B is a sentence, either Γ implies B or it doesn't. This presupposition seems reasonable to me (though it is not beyond question); but even if it is rejected, the set-theoretic definition of validity gives a useful first approximation, that can be grasped by the advocate of classical logic as a first step toward understanding how to reason in the non-classical logic.

What if we shift from explaining validity to explaining truth? In my view, model theory plays at best a very indirect role in explaining truth. Rather, truth is directly explained by means of Schema (T), and model theory enters in only in helping us

understand the logical connectives that occur in instances of Schema (T). More fully, (i) model theory gives an account of validity for the non-classical logic, which tells us a lot about how to reason with the connectives in the logic; (ii) once we come to understand how to reason in the logic we will fully understand its connectives; (iii) when we understand the connectives, together with the primitive non-logical symbols, then we will understand the sentences of the language; and (iv) that means that we will understand what it is for a sentence of the language to be true, given that we accept all instances of the schema (T).[30] So it is only through its role in explaining validity that model-theoretic semantics helps convey an understanding of truth for *L*-sentences.

I won't argue here for this positive view of how we understand the notion of truth for *L*-sentences. What I will do, though, is try to undermine the idea that a model theoretic semantics could have any more direct role to play in our understanding of truth.

The first point to be made here depends on the fact that the model theory is a model theory for a non-classical logic, but is being given within an effectively classical part of the language, namely set theory. The point is an obvious one: if we are to use a logic without excluded middle to handle the paradoxes, such instances of excluded middle as

$$True(\langle Q \rangle) \vee \neg True(\langle Q \rangle)$$

must be unacceptable (where Q is a Liar sentence). But if 'True' were defined in set-theoretic terms, we would have to accept it, given that excluded middle holds within set theory. So a model-theoretic semantics for a non-classical language can't possibly have the means for defining the notion of truth. (It also can't have the means for defining the notion of determinate truth or any other such notion; for according to G-solutions no such notion can be subject to excluded middle, as they would have to if defined in classical set theory. I will have much more to say about determinate truth in later sections.)

There is a second point with the same conclusion, and this one arises even for the model theory of classical languages. It is based on Tarski's theorem about the undefinability of 'true-in-*L*' in the 'true'-free portion of *L*. Tarski stated the theorem for classical languages only, but obviously it extends to a non-classical language *L* if we assume that its 'true'-free portion L_0 is classical, since a definition of 'true-in-*L*' in L_0 would yield a definition of 'true-in-L_0' in L_0.

[30] Of course, we want to be able to prove generalizations about truth that don't follow from the instances of (T) (though maybe they follow from the schema understood in a broader sense—see [7]). It is doubtful, though, that these are required for understanding the notion. Even if they are, my basic point is that our understanding of 'true' is given by the acquisition of a theory that contains it; the assumption that this theory consists only of Schema (T) is completely inessential to this basic point.

How is it that we can use classical set theory to define 'the semantic value of A relative to a model', but can't use it to define 'true'? The answer is that semantic value is relative to a model, and that truth (in the intended sense, the sense that obeys the Tarski schema (T)) is not. And the crucial point about the relativity to a model is that *in a model, the quantifiers are restricted to the members of a set; they do not range over absolutely everything*. Without this, the explicit definition of semantic value would not be possible.

If we want to think of the model-theoretic semantics as telling us not just about validity but about truth, then we will have a special interest in what we might call *homophonic models*: models which assign to a name its real bearer and analogously for function symbols, and which in the classical case assign to a predicate those objects in its real extension *that are also in the domain of the model*. But truth-in-a-homophonic-model must be distinguished from truth (even in the classical context where claims about truth obey the law of excluded middle). To say that a sentence is true in a homophonic model M is to say in effect that it would be true if its quantifiers were restricted to the domain of M. That can be defined (if the model M is definable); but by its very model-relativity it diverges from the notion of truth.

Consider a classical model for the 'true'-free part L_0 of the language L. L_0 includes standard set theory. Suppose we take a highly natural model for L_0, say the homophonic model whose domain consists of all non-sets together with all sets of rank less than the first inaccessible cardinal; call this homophonic model M_1.[31] This assumes of course that there are inaccessible cardinals; otherwise there would be no such model, so we'd have to use a different example. But now consider the sentence 'There are inaccessible cardinals': it's true, but false in M_1, i.e. has semantic value 0 in M_1; its negation is false, but has value 1 in M_1. Having semantic value 1 in M_1 doesn't correspond to truth, or to determinate truth, or anything like that, even in the classical sublanguage L_0 of L. The point made here for M_1 applies to any other model that can be defined within set theory, by Tarski's Theorem, and this includes all models of set theory that are at all 'natural' . (Indeed, the point has an extension to 'unnatural' models of set theory that are not set-theoretically definable: see [3], n. 24.)

The term 'model' is sometimes employed in a broader sense than I have been taking it, a sense in which we give a model by specifying a domain that needn't be a set. If this is done in the context of a theory in which we quantify over proper classes as well as sets, it obviously changes nothing important: the sentence 'There are proper classes' will come out having value 0 in the model even though it is (according to the theory) true (and even though the model is homophonic and quasi-correct). If

[31] M_1 is thus 'quasi-correct' (as defined at the end of Section 5) as well as homophonic. (These requirements are somewhat independent: e.g. quasi-correctness requires homophony in the set-theoretic vocabulary but not elsewhere.)

it is done in the context of a theory in which we don't quantify over proper classes but regard 'proper class' talk as a dispensable manner of speaking to be construed in terms of language, then Tarski's undefinability theorem applies: the notion of truth or semantic value *in a proper class model* is not explicitly definable in the set-theoretic language; in reasoning with it, we are going beyond standard set theory, we are reasoning in a set theory expanded by adding a notion of set-theoretic truth. (And of course we have then left the notion of truth for arbitrary sentences in this expanded language undefined.)

One might think I am making too much of the fact that 'true-in-L_0' isn't explicitly definable in L_0, when L_0 is a classical language: after all, it is inductively definable in L_0,[32] and the problem is only that this inductive definition can't be made explicit because the quantifiers range over everything. I'm not sure why the possibility of inductive definition in the classical case should be thought to undermine what I've said, but there's no need to go into that: for it is completely irrelevant to the case of actual interest in this chapter, the semantics of the non-classical languages used for the paradoxes. For in every case of which I am aware, the model-theoretic semantics used for those languages requires the quantifiers to be restricted *even in giving the inductive definition* of semantic value. This is certainly so for any model theoretic semantics that builds on a Kripke-like model theory (see [11]), for that requires an inductive construction *whose first step is an explicit definition of truth for the 'true'-free sublanguage*. If the quantifiers in the language weren't restricted to the members of a given set, the inductive specification couldn't get off the ground.

Indeed, for most G-solutions the point is even more striking. They build on Kripke's theory of truth, and thus are subject to the previous observation. But in addition, the cardinality of the evaluation space V in the usual semantics is larger than the cardinality of the starting model—see note 20 for why this is so.[33] This fact seems to me to seriously undermine the idea that we can somehow extrapolate an understanding of a model-independent notion like truth or determinate truth from the model-dependent notions. For if we were to try to somehow extrapolate

[32] Or to be pedantic, the term 'satisfies in L_0', from which 'true in L_0' is explicitly definable, can be inductively defined in L_0.

[33] The argument in n. 20 turned on the fact that V was assumed to be c-complete and its maximal element to be c-join irreducible. I imposed these requirements to ensure that every formula receive a value on every assignment function and that $\exists$-elimination be validated, but Philip Welch has pointed out to me that the assumptions are stronger than needed for these purposes: after all, the subspace V_0 of V consisting of those values that are actually assigned to formulas relative to at least one assignment function could be used instead, and it has only cardinality c. Use of this sparse subspace seems unnatural, but it would evade the argument of this paragraph of the text; though the argument of the previous paragraph would still remain. (On the unnaturalness: it's a bit like avoiding the use of the uncountable real number system in a theory by using instead a countable extension of the rationals in which only those bounded sets *that are definable in the language of the theory* have least upper bounds.)

from the case of models on domains with a cardinality c and valuation spaces with a cardinality $f(c)$ strictly bigger than c to the case of 'models' whose domain is absolutely everything, this would seem to require a valuation space 'strictly bigger than absolute infinity' , i.e. not only with more members than any set but with more members than 'the totality of absolutely everything'. I don't think such an extrapolation possible: model-theoretic notions are one thing, truth and determinate truth are something else again.

To summarize this section, it is very dangerous to draw conclusions about truth and related notions from model-theoretic semantics, for at least two reasons: (a) because the model-theoretic semantics is in a classical metalanguage, so that excluded middle is assumed throughout; (b) because even the homophonic models falsify how the language works, by taking the quantifiers to range over a certain set M (the domain of the model) rather than ranging over absolutely everything. Because of these two facts, the notion of semantic value is inevitably a somewhat artificial construction that can only be understood as model-relative, and drawing conclusions about how sentences are to be evaluated with respect to properties that are *not* model-relative (for instance, truth, determinate truth, and so forth) is highly problematic. If such conclusions can be drawn at all, it is only with extreme caution.

9 The simplest revenge arguments

I think that the points in the preceding section can be used to undermine the following two revenge arguments. (More difficult revenge arguments will be considered later on.) These two arguments don't depend much on the details of the G-semantics.

Simple inferential revenge argument: According to the semantics, the space of semantic values is partitioned into two classes, the designated and the undesignated; and *the semantics assumes designatedness to be a classical notion*, that is, each sentence is either designated or undesignated. But suppose we had *in the language* the predicate 'has a designated value'. The predicate would not only obey excluded middle, it would also need to be 'truth-like' in the sense of Section 3. That is, the following inferences would need to be valid:

[*Des*-Elim] $\quad Designated(\langle A\rangle) \vDash A$

and

[*Des*-Incoher] $\quad A, \neg Designated(\langle A\rangle) \vDash \bot$ (where $\bot$ is an absurdity).

Indeed, we'd probably want to strengthen the former to

$$\vDash Designated(\langle A\rangle) \rightarrow A$$

and the latter to

$$A \vDash Designated(\langle A\rangle).$$

But even without these strengthenings, [*Des*-Elim] and [*Des*-Incoher] lead to absurdity, using a 'super-Liar' sentence Q_* that asserts that it doesn't have a designated value, by the reasoning

early in Section 3. (To review: $Designated(\langle Q_* \rangle)$ and $\neg Designated(\langle Q_* \rangle)$ each imply Q_* (the first by [*Des*-Elim], the second by the definition of Q_*), so using reasoning by cases plus excluded middle we get Q_*; but then by the definition of Q_* we also get $\neg Designated(\langle Q_* \rangle)$, and these two claims together are absurd by (*Des*-Incoher).)

It seems that if we allow the notion of designatedness into the language, assumptions about it that appear almost inevitable lead to contradiction.

Before evaluating this argument, let's consider a semantic variant:

Simple semantic revenge argument: As in the inferential version, we argue that if we had in the language the predicate 'has a designated value' then we could form a 'super-Liar' sentence Q_*; so

(*) $Q_* \leftrightarrow \neg Designated(\langle Q_* \rangle)$

must have a designated value. But this requires

(**) Q_* has designated value if and only if $\neg Designated(\langle Q_* \rangle)$ has designated value.

But if Q_* has a designated value, then $Designated(\langle Q_* \rangle)$ should too, and so $\neg Designated(\langle Q_* \rangle)$ should not have designated value. So given (**), the assumption that Q_* has designated value is absurd. Similarly, if Q_* does not have designated value, $\neg Designated(\langle Q_* \rangle)$ should have designated value; so given (**), the assumption that Q_* does not have designated value is also absurd. But Q_* either has a designated value or doesn't (the semantics being classical), so we are landed in an absurdity either way.

Again, it looks like something really unattractive is required, if we allow the notion of designatedness into the language.

What the proponent of either form of the argument holds, then, is that if we allow a designatedness predicate into the language that we are semantically evaluating, we have a paradox: we are led into contradiction unless we abandon an assumption that is central to the intuitive meaning of the notion. And if we don't allow such a predicate into the language that we are semantically evaluating, then our solution to the paradoxes works only because the language being evaluated is expressively incomplete.

A *possible* reply to both arguments, though not one I find at all attractive, is in terms of 'levels of language'. According to this possible reply, we have to introduce a whole hierarchy of value spaces $V_{(1)}$, $V_{(2)}$, ... and a corresponding hierarchy of designatedness predicates; the paradoxes arise, on this view, by assuming that sentences containing a given designatedness predicate Des_α are themselves to be evaluated in terms of Des_α, when in reality they are to be evaluated in terms of $Des_{\alpha+1}$. (In the semantic version, the claim would presumably be that sentences containing Des_α are simply unevaluable in $V_{(\alpha)}$, but only in $V_{(\alpha+1)}$ and higher. In the inferential version, the claim would presumably be that the inference rule $A \vDash Des_\alpha(\langle A \rangle)$ (or $A, \neg Des_\alpha(\langle A \rangle) \vDash \bot$) must be restricted to the case where A has no *Des* predicate subscripted α or higher.) These solutions are

formally adequate, but in invoking such a hierarchy of value spaces and of unrelated Des_α predicates they seem completely outside the spirit of G-solutions to the paradoxes.

Fortunately, such a 'level of languages' approach is completely unnecessary. The proper reply to the simple arguments, I think, is that for the reasons discussed in the previous section, any intelligible 1-placed predicate of having designated value is model-relative. Let's stick to a predicate that relativizes to a *specific* model: 'is designated in the valuation based on M_0'. (It's easy to see that predicates like 'is designated in *all* valuations based on models of such and such sort' and 'is designated in *some* valuation based on a model of such and such sort' could only make things worse.) Now for these relativized predicates, there is no paradox: they are already in the language L, they obey excluded middle, they fail to obey the assumptions [Des - Elim] and [Des - Incoher] used in the two arguments, *but that failure is in no way surprising or paradoxical precisely because of the relativized nature of the predicate*. For example, suppose the model M_0 is the homophonic model whose domain consists of those objects of rank less than the first inaccessible cardinal. Then 'There are inaccessible cardinals' is undesignated relative to that model, even though it is (determinately) true; and its negation is designated relative to that model, but (determinately) false. In short, if 'designated' is interpreted as 'designated relative to M_0' then lots of perfectly ordinary sentences (sentences of set theory, *not containing 'true' or other suspect terms*) have precisely the 'paradoxical' features of the sentence that asserts its own lack of 'M_0-designatedness' . There simply is no paradox here.

Obviously the proponent of the simple revenge problem doesn't intend 'designated' to be understood as model-relative. The question then arises, how is it to be understood? I do not deny that it is possible to introduce into the language an operator (which I prefer to call 'determinately') with many of the features that the proponent of revenge wants, and which is *not* model-relative. Indeed, as we'll see in the next section, such predicates are already definable in L! But such predicates only breed paradox if they satisfy all the assumptions used in the derivations above. It turns out that one can get predicates that satisfy *most* of the assumptions used in the derivations above; the one place they fail is that excluded middle cannot be assumed for them. So there is a revenge problem (of the sort considered in this section) only if there is reason to think that we can understand a notion of 'designatedness' that obeys those other assumptions *plus excluded middle*.

And why assume that? I think what underlies the simple revenge problem is the thought that the model-relative designatedness predicates all obey excluded middle, so there must be an absolute designatedness predicate that does too. But this assumption seems to me completely unwarranted: one just can't assume that one can extrapolate in this way from the case of model-relative predicates, which make sense only by virtue of 'misinterpreting' the quantifiers as having

restricted range, to the unrelativized case where no such 'misinterpretation' is in force. In the case of G-solutions, even the choice of value-space depends on the initial restriction of domain; if one tries to idealize away the restriction of domain, one is left without a choice of value-space. How one is supposed to be left with an intuitive understanding of an absolute notion of designatedness (even one that can only be formulated in a richer language) is beyond my comprehension.

Part Three Determinacy and Iterated Determinacy

10 Determinacy and 'strengthened and weakened liar sentences'

Once we assume a G-semantics, there is no danger that there are any genuine paradoxes statable in the language *L*: any apparent paradox statable in the language has a solution that is consistent with all instances of (T) and (S). If there is a revenge worry, it is that the language is expressively weak, and that there are concepts we need that if added to the language would breed new paradoxes. For instance, the Liar sentence is clearly somehow 'defective' , but we've seen that we can't explain its defectiveness as its being neither true nor false; can we explain this in some other way? It's natural to say that its defectiveness consists of its being neither *determinately* true nor *determinately* false. We can take 'determinately' to be an operator *D* taking formulas to formulas; from that we can form predicates of determinate truth and determinate falsity, viz., $D[True(x)]$ and $D[True(neg(x))]$ (and analogously, predicates of determinate satisfaction).

A worry is that if we add such a determinacy operator to the language (in a way that allows us to say of the Liar sentence that it is neither determinately true nor determinately false), we will inevitably be led to new paradoxes that cannot be consistently treated. A weaker worry is that the motivations for introducing a notion of determinacy will eventually lead to a hierarchy of richer and richer languages. I will argue that both worries are unfounded.

We want to allow the operator *D* to apply to formulas that themselves contain *D*. (This is a precondition to avoiding a hierarchy of languages.) The key to avoiding paradox is that excluded middle won't be assumed for claims of determinate truth: we *don't* want that every sentence is either determinately true or not determinately true. (Analogously for determinate satisfaction.) Because of this and the fact that the semantics is given in classical logic, we can't straightforwardly define the notion of determinate truth in terms of the semantics.

Instead, let's impose some conditions that a reasonable determinacy operator should satisfy. From a model-theoretic viewpoint it seems quite reasonable to assume

that D is value-functional and satisfies the following:

(a) If $|A|_{M,s} \leq |B|_{M,s}$ then $|DA|_{M,s} \leq |DB|_{M,s}$.
(b) If $|A|_{M,s} = 1$ then $|DA|_{M,s} = 1$
(c_w) $|DA|_{M,s} \leq |A|_{M,s}$

and probably the strengthened form of that

(c) If $0<|A|_{M,s}<1$ then $|DA|_{M,s}<|A|_{M,s}$, and if $|A|_{M,s} = 0$ then $|DA|_{M,s} = 0$.

(a), (b), and (c_w) correspond to natural inferential principles: (a) to the principle $A \rightarrow B \vDash DA \rightarrow DB$, (b) to $A \vDash DA$, and (c_w) to $\vDash DA \rightarrow A$. The inferential content of the remainder of (c) is that $A \rightarrow DA \vDash A \vee \neg A$. (The converse inference $A \vee \neg A \vDash A \rightarrow DA$ follows from (b).)[34]

Conditions (a), (b), and (c_w) are clearly insufficient for D to count as a determinacy operator, for they are compatible with D being the identity operator. (c) partially rectifies that, but is insufficient to guarantee that we can legitimately declare the Liar sentence Q to be not determinately true (or indeed, to guarantee that we can declare $\neg DA$ for any A for which we cannot declare $\neg A$). Model theoretically, what we need is that DQ have value 0. For that, the following is sufficient (and in most versions of G-semantics, also necessary):

(d) If $|A|_{M,s} \leq |\neg A|_{M,s}$ then $|DA|_{M,s} = 0$,

which corresponds to the 'modified *reductio* principle' $A \rightarrow \neg A \vDash \neg DA$.[35]

If one takes the model theory sufficiently seriously one may want to replace (a)–(d) by a slightly stronger condition:

(e) There is an operator ∂ on the space V_c of values such that if $|A|_{M,s}$ is v, $|DA|_{M,s}$ is ∂v, and which satisfies analogues of (a)-(d). That is, which satisfies

(∂a) If $v_1 \leq v_2$ then $\partial v_1 \leq \partial v_2$
(∂b) $\partial 1 = 1$
(∂c) $\partial 0 = 0$, and if $0<v<1$, then $\partial v<v$
(∂d) If $v \leq v^*$ then $\partial v = 0$, where * is the operator that corresponds to negation.

[34] It doesn't follow that $A \rightarrow DA$ is fully equivalent to excluded middle, i.e. that $|A \rightarrow DA| = |A \vee \neg A|$, and in typical G-logics (e.g. those meeting condition (III_s) of note 23) this can fail: for instance, where Q is the Liar sentence, $Q \rightarrow DQ$ may have value 0, which can never be the case for an instance of excluded middle.

[35] **Sufficiency:** Since $|Q| = |\neg True(\langle Q \rangle)| = |\neg Q|$, (d) requires that $|DQ| = 0$. **Necessity:** If $|A|_s \leq |\neg A|_s$ then $|A|_s \leq |A \wedge \neg A|_s$. In every G-semantics I know of (e.g. that of [5]), we have that for any A, $|A \wedge \neg A|_s \leq |Q|$. Assuming that, the above yields that if $|A|_s \leq |\neg A|_s$ then $|A|_s \leq |Q|$. But then (a) yields that if $|A|_s \leq |\neg A|_s$ then $|DA|_s \leq |DQ| = 0$.

This is stronger, because V_c might contain values that no sentence of the language could possess; the most that follows from (a)-(d) is that a *partial* operator ∂ *defined on the subset of* V_c *that is in the range of the assignment function* satisfies (a)-(d), but (e) adds that this operator is total. The stronger condition seems natural if not irresistible, and I have no objection to adding (e) as a requirement. Call an operator D satisfying these conditions a *determinacy operator* (and call the corresponding ∂ satisfying (∂a)-(∂d) a ∂*-operator*). If we require only (c_w) in addition to (a), (b), and (d), D will be called a *weak determinacy operator*; similarly, weakening (∂c) to

$$(\partial \mathbf{c}_w) \qquad \partial v \leq v$$

gives the requirements on a *weak* ∂*-operator*.

Given a G-semantics for a language without a determinacy operator, can we extend it to one with such an operator? We certainly can extend it to one with a weak determinacy operator; and for all versions of G-semantics I know of, we can extend it to one with a full determinacy operator. The simplest way to show this is to explicitly define such an operator D from the connectives we already have in the language. One particular such D (I'll call it **D**) can be defined as

$$A \wedge \neg(A \rightarrow \neg A).^{36}$$

Obviously it corresponds to an operator in the underlying space, viz. $\boldsymbol{\partial} v =_{df} glb\{v, (v \Rightarrow v^*)^*\}$. That the conditions ($\partial$a), ($\partial$b), ($\partial c_w$), and ($\partial$d) are satisfied is apparent: e.g. to verify (∂d), we simply observe that if $|A|_s \leq |\neg A|_s$ then $|A \rightarrow \neg A|_s = 1$, so $|\neg(A \rightarrow \neg A)|_s = 0$, so $|A \wedge \neg(A \rightarrow \neg A)|_s = 0$. And the full ($\partial$c) is satisfied in the versions of G-semantics I'm familiar with, e.g. that of [5]. I'm mostly interested in G-semantics in which full determinacy operators are definable, but weak determinacy operators would in fact fit my main purposes, and I will rely below on the fact that they are definable in *every* G-semantics.

An advantage of treating the determinacy operator as defined within the original language is that doing so settles the application of 'True' and 'Satisfies' to sentences containing this operator, and settles it in such a way that the biconditionals (T) and (S) and the corresponding intersubstitutivity theses are bound to hold even for sentences that involve the determinacy operator. Thus with 'determinately' defined as **D**, it is immediate that **there can be no new paradoxes of determinacy** .

Two examples that might initially be thought paradoxical are the 'strengthened Liar sentence' Q_{-1} which directly or indirectly asserts that it is determinately not true (and hence must be a fixed point of the 'strong negation operator' **D**¬), and

[36] In [3] I used the alternative definition $A \wedge [T \rightarrow A]$, where T is a logical truth; this is equivalent to the definition in the text in the particular G-logic considered there, but I believe that in G-logics for which the equivalence fails, the operator in the text is more useful.

the 'weakened Liar sentence' Q_1 which directly or indirectly asserts that it is not determinately true (and hence must be a fixed point of the 'weak negation operator' $\neg\mathbf{D}$). Let's focus on the latter. Q_1 is equivalent to $\neg\mathbf{D}[\mathit{True}(\langle Q_1\rangle)]$, and hence (by the Tarski biconditionals and substitutivity principles) to $\neg\mathbf{D}Q_1$, so

$$(^*) \quad |Q_1| = |\neg\mathbf{D}Q_1| \text{ (or equivalently, } |\neg Q_1| = |\mathbf{D}Q_1|).$$

What can we say about the value of $\mathbf{D}Q_1$? First, $|\mathbf{D}Q_1|$ isn't 0: for if it were, then by $(^*)$, $|Q_1|$ would be 1, and we'd have a gross violation of (b). Second, $|\mathbf{D}Q_1|$ is strictly less than $|Q_1|$: for since $|\mathbf{D}Q_1| \neq 0$, we have by condition (d) that $|Q_1| \not\leq |\neg Q_1|$, so by $(^*)|Q_1| \not\leq |\mathbf{D}Q_1|$; together with $|\mathbf{D}Q_1| \leq |Q_1|$ (condition (c_w)), this yields $|\mathbf{D}Q_1| < |Q_1|$. So application of the operator corresponding to **D** strictly lowers the value of Q_1, but doesn't reduce it to 0.

We do have, though, that a double application of this operator reduces the value to 0; that is, $|\mathbf{DD}Q_1| = 0$ (from which it follows that *it is not determinate that Q_1 is not defective*, i.e. $\neg\mathbf{D}\neg[\neg\mathbf{D}Q_1 \wedge \neg\mathbf{D}\neg Q_1]$[37]). The argument for $|\mathbf{DD}Q_1| = 0$ is that by condition (c_w), $|\mathbf{D}Q_1| \leq |Q_1|$; so by $(^\star)$, $|\mathbf{D}Q_1| \leq |\neg\mathbf{D}Q_1|$; and $|\mathbf{DD}Q_1| = 0$ then follows by condition (d). It follows, of course, that **DD** (the result of applying **D** twice) is strictly stronger than **D**. (The same conclusion would have emerged from studying the 'strengthened Liar' : the fact that **DD** is strictly stronger than **D** shows up in the fact that $|\mathbf{D}\neg Q_{-1}| > 0$ but $|\mathbf{DD}\neg Q_{-1}| = 0$.)

Notice that this argument doesn't turn on the specific definition proposed for **D**; it doesn't even turn on the fact that **D** is definable in terms of the other connectives. Rather, it turns only on the general features of a G-semantics plus conditions (b), (c_w), and (d); not only was (c) weakened to (c_w), but (a) wasn't used. Indeed, we didn't even use the full strength of (b), we only used

$$(\mathbf{b}_w) \text{ If } |A|_{M,s} = 1 \text{ then } |DA|_{M,s} > 0,$$

(whose inferential analog is $A, \neg DA \vDash \bot$, where $\bot$ is an absurdity). (b_w), (c_w), and (d) are the conditions for what I'll call a *very weak determinately operator*. We see then that for any such operator in a language with a G-semantics, *DD* is strictly stronger than *D*. Calling an operator *D idempotent* if *DD* is the same as *D*, we get

Conclusion : No operator in a language with a G-semantics that meets the conditions for being even a *very weak* determinately operator can be idempotent.[38]

[37] For this simplifies to $\neg\mathbf{D}[\mathbf{D}Q_1 \vee \mathbf{D}\neg Q_1]$, which by the nature of Q_1 is equivalent to $\neg\mathbf{D}[\mathbf{D}Q_1 \vee \mathbf{DD}Q_1]$, which by two successive applications of $\neg\mathbf{DD}Q_1$ is valid.

We can't assert that Q_1 is defective (or, of course, that it isn't); only that it is 'possibly' defective, i.e. not determinately non-defective.

[38] Also, if *D* meets those conditions then $D\neg D$ is strictly stronger than $\neg D$: for $|D\neg DQ_1|$ is $|DQ_1|$ and $|\neg DQ_1|$ is $|Q_1|$, and we've already seen that $|DQ_1| < |Q_1|$.

(It was implicit in our result on 'truth-like operators' in Section 3 that no very weak determinately operator in a G-semantics can be bivalent, i.e. can be such that the value of any sentence of form DA is always 0 or 1; indeed, that result doesn't even require (d). It's also worth noting that for any operator that meets the full condition (c), idempotence immediately implies bivalence: (c) implies that for any A, if $|DDA|_s = |DA|_s$ as demanded by idempotence, then $|DA|_s$ can only be 1 or 0. So (b_w) and the full (c) immediately rule out idempotence in any G-semantics.[39])

I must emphasize that these conclusions only apply to operators in languages with a full-fledged G-semantics; that is, languages with a DMC-semantics *for which the Tarski biconditionals hold*. Obviously if one gives up the latter assumption, the conclusion no longer need hold. For instance, suppose we were to stipulate that the original language is to be extended by the addition of an operator D^* which is stipulated to behave in the semantics as follows:

$$|D^*A|_{M,s} \text{ is 0 if } |A|_{M,s} < 1, \text{ and } |D^*A|_{M,s} \text{ is 1 if } |A|_{M,s} \text{ is 1.}$$

(D^* is of course just the operator version of the alleged 'absolute designateness' predicate discussed in the previous section.) Such a D^* is by stipulation bivalent, and is easily seen to satisfy all of the conditions (a)-(d); our conclusions thus imply that adding it to the language would force a failure of the Tarski biconditionals, which we knew already (from the discussion of attempts at a model-theoretic revenge argument).

Might we contemplate settling for a determinacy operator that maintains idempotence and the Tarski biconditionals by weakening one of the conditions (b_w), (c_w), and (d)? One's first inclination might be to weaken (d), but as remarked in note 35, there is no way to do this in the best-known versions of G-semantics without either weakening (a) or giving up the idea that $|DQ| = 0$, either of which seems an intolerably high cost to pay. (And even in a G-semantics where one could weaken (d) without incurring these costs, keeping idempotence together with (b_w) would rule out acceptance of the full (c).)

Another way to keep idempotence in a G-semantics would be to say that the D operator has non-trivial application only to formulas that don't themselves contain D: the line would be that if A contains D, $|DA|_s = 0$, so DD is the 0 operator and hence trivially idempotent.[40] This would allow us to preserve (c) and (d), and it generates a restriction on (b_w) that blocks the above proof (and it generates a restriction on

[39] In inferential form, (c) amounts to the law $B \rightarrow DB \vDash B \vee \neg B$, which implies $DB \rightarrow DDB \vDash DB \vee \neg DB$. Idempotence would establish the premise, so we'd get excluded middle for D claims, which would suffice for inferential paradox.

[40] Of course, this will have to affect the application of the D operator to some sentences that don't contain 'D' but contain 'True' or 'Satisfies', if the schemas (T) and (S) are to be preserved.

(a) as well): in particular, the various 'determinate Liar' sentences Q_α ($\alpha \geq 1$) now all get value 1. But limiting the scope of the determinacy operator in this way seems to me a very high cost. One could ameliorate that cost by introducing a hierarchy of determinacy operators, e.g. an operator D_2 that applies non-trivially to formulas containing the 'ground level' determinacy operator D; but D_2 couldn't be explained as DD since DD would just be D. It seems to me that introducing such a hierarchy of operators would throw away whatever virtues idempotence might be thought to have, and that the idea of a hierarchy of *primitive* determinacy operators has far less appeal than a hierarchy obtained by iterating a single determinacy operator.

The fact that the G-semantics rules out the existence of an idempotent operator satisfying the conditions of a determinacy operator (or even a *very weak* determinacy operator) would lead to a revenge problem if we had reason to believe that such an operator was intelligible: it would show a serious expressive limitation in any such language. But I maintain that there is no good reason to think that idempotent determinacy operators, or even idempotent very weak determinacy operators, are intelligible. I've already considered one argument for the intelligibility of the particular idempotent operator D^*, and argued that this argument rests on a misunderstanding of the significance of the model theory. But the possibility that there are more sophisticated arguments for idempotence remains to be considered.

11 Iterations of determinacy operators

The impossibility of idempotent operators in a G-semantics is a more stringent ban that might at first appear. To see this, observe that if D is any determinacy operator then so is DD (the result of applying it twice); that is, DD satisfies (a)-(d) if D does. Similarly, if D is any *weak* determinacy operator then so is DD. (Not so for *very* weak.) Because of this, our conclusion that if D is a weak determinacy operator in a G-semantics then it can't be idempotent can be extended: not only can't D be idempotent, DD can't either. That is, DD can't be identical to $DDDD$; from which it easily follows that it can't be identical to DDD. Continuing in this vein, we can argue that as n increases, the result D^n of applying D n times becomes strictly stronger as n increases.

We can also extend the iteration process a good way into the transfinite, as I will discuss in detail in Part Four. The basic idea, restricted for simplicity to the case where A is a sentence, is that for each limit ordinal λ for which the iteration is defined, $D^\lambda A$ says that for all $\alpha < \lambda$, $D^\alpha A$ is true; as a result,

$$|D^\lambda A| \text{ is the greatest lower bound of } \{|D^\alpha A| \mid \alpha < \lambda\}.$$

More generally, we can get that even when A can contain free variables,

$$\text{(LIM)} \quad |D^\lambda A|_s \text{ is the greatest lower bound of } \{|D^\alpha A|_s \mid \alpha < \lambda\}.$$

It's then easy to verify that as long as the iteration process is defined in accordance with (LIM) at each stage, it always leads from determinacy operators to determinacy operators and from weak determinacy operators to weak determinacy operators. So the anti-idempotence result shows that **we get a hierarchy of operators that become strictly stronger for as long as the iteration is satisfactorily definable** (i.e. definable in a way that accords with (LIM)).[41] The limits on how far it is so definable raise interesting philosophical issues which I will discuss in Part Four.

This fact that the hierarchy never collapses to idempotence can be directly checked, by producing a hierarchy of 'increasingly paradoxical' sentences. The simplest such hierarchies are the **transfinite Liar hierarchies**. There are two of them, 'going in opposite directions' ; I will focus on the fixed points Q_α of the ' increasingly weak negations' $\neg, \neg D, \neg DD, \ldots$, but I could just as well have used the fixed points $Q_{-\alpha}$ of the 'increasingly strong negations' $\neg, D\neg, DD\neg, \ldots$.[42] So: Suppose we have defined D^α for some ordinal α. Then we can find an 'α-level weakened Liar sentence' Q_α which says $\neg D^\alpha True(\langle Q_\alpha \rangle)$. By the Tarski biconditionals and substitutivity principles (which by definition hold in any G-semantics), it follows that

$$(^{**}) \quad |Q_\alpha| = |\neg D^\alpha Q_\alpha|.$$

$|D^\alpha Q_\alpha|$ isn't 0: for if it were, then by (**), $|Q_\alpha|$ would be 1, and we can easily see (using (b) and (LIM)) that then $|D^\sigma Q_\alpha|$ would have to be 1 for all σ, contradicting our assumption. However, $|D^\alpha Q_\alpha| \leq |Q_\alpha|$ (by (c_w)); so by (**) $|D^\alpha Q_\alpha| \leq |\neg D^\alpha Q_\alpha|$; so by (d), $|D^{\alpha+1} Q_\alpha| = 0$. So $D^{\alpha+1}$ is a strictly stronger operator than D^α (from which it follows that it is not $D^{\alpha+1}$-true that Q_α is non-defective[43]).

To summarize, the situation is that for a fixed Q_α, the $D^\sigma Q_\alpha$ get stronger and stronger as σ increases, until σ reaches $\alpha + 1$; after that point no strengthening of

[41] Vann McGee and Elia Zardini have called my attention to an argument, informally suggested by Timothy Williamson [23], that idempotence is bound to be reached by stage ω. But the argument turns on the assumption that D distributes over $\forall$ (in the conditional form of this assumption: that is, $\forall n DA(n) \rightarrow D\forall n A(n)$, not merely the unproblematic rule form $\forall n DA(n) \models D\forall n A(n)$). D-distributivity (in this conditional form) is not valid in G-semantics. (The special case where the quantification is restricted to natural numbers and $|A(n)|$ is decreasing in value as n increases would suffice for Williamson's argument, but distributivity fails even in that special case.) Indeed, distributivity of D over finite conjunctions fails (though it holds when the conjuncts have values that are $<$-comparable).

[42] Each $Q_{-\alpha}$ behaves very much like the negation of the corresponding Q_α.

[43] Indeed, for finite α, the '$\alpha + 1$' can be replaced by 'α' in the parenthetical claim. The proof of the parenthetical claim is a generalization of that in n. 37: to say that it is not D^σ-true that Q_α is non-defective is to say that $\neg D^\sigma \neg[\neg DQ_\alpha \wedge \neg D\neg Q_\alpha]$, i.e. $\neg D^\sigma [DQ_\alpha \vee D\neg Q_\alpha]$, i.e. $\neg D^\sigma [DQ_\alpha \vee DD^\alpha Q_\alpha]$. This reduces to $\neg D^{1+\sigma} Q_\alpha$ (using the fact that $\neg D^{\alpha+1} Q_\alpha$), and this is valid whenever $1 + \sigma \geq \alpha + 1$; i.e. when $\sigma \geq \alpha + 1$ and α is infinite, or $\sigma \geq \alpha$ and α is finite.

the sentence by adding 'determinately' is possible, since the sentence already has value 0. But there are other sentences, e.g. the Q_β for $\beta > \alpha$, for which the iteration can proceed farther before collapsing to value 0; so the operators D^σ *never* collapse to the operator D^* or to any other idempotent operator.

The claim that the determinacy operator D, and the corresponding operator ∂ on the space of values, never collapses into idempotence may seem as if it couldn't be true: after all, for any partially ordered set V_c there is a cardinal d greater than the cardinality of all V_c-chains (linearly ordered subsets of V_c). But then for any formula A and assignment function s, the chain $\{|D^\alpha A|_s\}$ has stopped decreasing prior to the initial ordinal for d; so letting ξ be that initial ordinal, it seems that D^ξ must be idempotent.

The problem with this argument is that the iteration of the D operator breaks down (becomes undefined) before reaching cardinality d. There are two reasons why this is so.

The first reason for the breakdown depends on the nature of each value space V_c (used for models of cardinality no greater than c). In order to ensure that quantified formulas get values in the space, I required that V_c be c-complete: that is, each subset of V_c of cardinality no greater than c must contain a least upper bound (or equivalently, each such subset must contain a greatest lower bound).[44] I also gave reasons (note 20) for expecting that the space would not contain least upper bounds (or equivalently, greatest lower bounds) for all sets of higher cardinality than c. But the only acceptable way to define ∂^λ for a limit ordinal λ would be to let $\partial^\lambda(v)$ be the greatest lower bound of all $\partial^\alpha(v)$ for $\alpha < \lambda$; so if λ has cardinality greater than c, there is no reason to think that ∂^λ is defined.[45] Indeed, the fact that we have shown that no iteration of ∂ is idempotent, together with the result of two paragraphs back that if ∂^ξ were defined it would be idempotent, shows that there must be subsets of V_c of cardinality no greater than that of ξ that have no greatest lower bounds.

But there is a second reason for the breakdown in the iteration of the D-operator, which occurs much earlier and is of more interest for generating revenge problems. It arises from the fact that the language L (like all languages, in any but a special technical sense of 'language') can contain only countably many expressions. From this it follows that there are *countable* α for which there is no operator D^α in the language; this is a much earlier breakdown in the iteration process than the one required by the

[44] The greatest lower bound of S is $(\sqcup\{v^* | v \in S\})^*$, where * is the operator on V corresponding to negation.

[45] Well, we could define $\partial^\lambda(v)$ as the greatest lower bound of all $\partial^\alpha(v)$ *for* α *in a sequence that is cofinal with* λ; but then if λ has no cofinal sequence of cardinality less than or equal to c, there is no reason to think that ∂^λ is defined. And the first ordinal of cardinality greater than c has no cofinal sequence of cardinality less than or equal to c, so this liberalization takes us no further.

previous paragraph. This earlier breakdown is a source of revenge worries, and in Part Four I will investigate in detail how it occurs.

So for both of these reasons, the iteration process breaks down, and that is how the claim that it never collapses to idempotence avoids absurdity. And to repeat, the no-collapse claim not only avoids absurdity, we have proved that it *must* hold given the basic assumptions of G-semantics and the very weak conditions (b_w), (c_w), and (d). Of course, one might hold on to idempotence and those weak conditions (or their inferential versions), if one were to give up the conditions on a G-semantics: for instance, if one were to give up the truth schema. But that has high costs. The main goal of the rest of the chapter is to argue that giving up on idempotence does not have an intolerably high cost, and indeed is very natural.

12 Non-idempotent determinacy versus stratified truth

One might think that a hierarchy of non-idempotent iterations D^α of a determinacy operator D would give rise to all the problems of stratified theories of truth and satisfaction in classical logic. But there are at least three reasons why this is not so (of which I take the first and third to be most important).

The first is that determinacy is a much more peripheral notion than truth, and we do have a unified notion of truth (and of satisfaction too). It is the notions of truth and satisfaction, not of determinate truth, that we need to use as devices of generalization. We've seen that stratifying the truth predicate would seriously cripple our ability to make generalizations; not so for a 'stratification' of the notion of determinacy.

The second point to make is that there is a serious disanalogy between the stratification of truth in classical theories and the 'stratification' of determinate truth here. For in this theory, all the determinacy predicates are defined by iterating a single determinacy operator (and using the truth predicate); whereas in the case of classical stratified truth theories, each truth predicate must be introduced separately.

The third point is that we can reasonably hope, for each α, that our overall theory of truth and satisfaction and determinacy is D^α *True*. This is in marked contrast to the classical truth theory case, according to which e.g. '$True_\alpha(\langle A\rangle) \rightarrow A$' is an important part of the theory but not true_α (but only $\text{true}_{\alpha+1}$). Thus the main objection that I raised against stratified theories (the one with which I closed Section 2) simply doesn't arise against the current theory.

Despite these three points, it may seem counterintuitive to suppose that there is no intelligible ' hyper-determinately' operator that is equivalent to '*True* and *DTrue* and D^2*True* and ...' (through all iterations of D); and that is what my account implies. Indeed it may seem not merely *counterintuitive*, it may seem *incompatible with point*

one, i.e. incompatible with the point that we can use 'True' to make generalizations that one couldn't make otherwise. These two qualms (counterintuitiveness and incompatibility with point one) are connected: one can't fully remove the intuitive pull of the first qualm without coming to understand why the second qualm is incorrect. And showing that the second qualm is incorrect requires the more precise treatment of the hierarchies that is to be given in Part Four. Even so, it is worth making a preliminary remark about the second qualm, and then addressing the first.

The incompatibility qualm: As I have been discussing the determinacy operators *so far*, the ordinal superscript 'α' in 'D^{α}' is not a variable; the intent has been rather that for each ordinal α in a suitable segment of the ordinals, there is a corresponding operator 'D^{α}' in the language. But (1) it looks as if we will be able to use the truth predicate to *get the effect of* quantifying over the α, even if the superscript isn't a variable. In particular, it looks as if we can express such thoughts as that for each α, the result of prefixing A with the α^{th} iteration of 'D' is true. But (2) that would be in effect just the application to A of an operator 'D^{α} *True* for all α' , and that looks like an idempotent determinateness operator that is the infinite conjunction of all the D^{α} *True*. I've argued against there being such an operator in the language (and will be arguing that such an operator is unintelligible); so what gives?

The long answer to this question will be given in Part Four. The very short answer is that we can indeed introduce a hierarchy of iterations D^{α} of D *for variable* α, and hence allow quantification over the α; however, in order to make the quantification well-behaved we must impose a bound on the α, and there are compelling reasons why there can be no unique such choice of bound.[46] Any bound we impose to keep the hierarchy of operators well-behaved can be relaxed, so that there is no maximal good bound. *Given any reasonable choice of bound for a hierarchy of iterations* D^{α} *of* D, we can then use the truth predicate to achieve what is in effect a quantification over the α, just as in the previous paragraph; but since the α are bounded, this will simply be another iteration of D in an enlarged hierarchy, so it does not achieve the intended effect.

The counterintuitiveness qualm is that it just seems as if we have a unified notion of hyper-determinate truth ('determinate truth in every reasonable sense of that term') corresponding to '*True* and *DTrue* and D^2*True* and . . . ' . Or if you like, a unified

[46] A slightly longer answer is that we can introduce a hierarchy of pseudo-iterations D^{α} of D for variable α, which we can quantify over unrestrictedly; but for large α these can't be viewed as genuine iterations of D, nor will they be determinacy operators in any reasonable sense; and there is no satisfactory way to restrict to 'good' ordinals (for which the D^{α} behaves as it is supposed to) except by restricting too far. (A still longer answer involves the idea that the notion of a 'good iteration' is a 'fuzzy notion' for which classical laws fail; this is why it is impossible to achieve satisfactory results by quantifying over only the good iterations.)

notion of 'defective in some reasonable sense of that term' , viz. '($\neg$*True* and $\neg$*False*) or ($\neg$*DTrue* and $\neg$*DFalse*) or ($\neg D^2$*True* and $\neg D^2$*False*) or . . . ' .[47]

I don't want to deny that we have these notions; but not every notion we have is ultimately intelligible when examined closely. A large part of the response to the counterintuitiveness qualm will be an argument, in Part Four, that the notion of 'the' hierarchy of iterations of *D* has a kind of inherent vagueness that casts doubt on there being a well-behaved notion of 'D^α-true for every α' ; and without that there is no reason to suppose that there is a well-behaved notion of 'determinately true in every reasonable sense of that term' . The apparent clarity of such notions is an illusion. (One can, to be sure, give *ill-behaved* definitions, that would seem well-behaved *if the inherent vagueness in the hierarchy were not taken into account*; so part of the response to the qualm will be to show how those definitions come to grief.)

A unified notion of hyperdeterminate truth, then, is basically something we should abandon. For this recommendation to be acceptable, following it had better not have the high intuitive costs that stratified truth theories have. And it doesn't. In addition to the three points made earlier in this section, I note the following. A substantial part of the counterintuitiveness of stratified truth theories stems from the fact that if such a theory were actually in use, each person would be under constant pressure to employ very high ordinal subscripts in order to ensure that what they said had sufficient strength; and because of ignorance about the subscripts employed by others whose views we are discussing, we would often end up employing too low a subscript to capture what we wanted to say. This is a point well illustrated by Kripke's discussion of Nixon and Dean (mentioned at the end of Section 2); one of the symptoms is that if Nixon says 'Everything Dean says is untrue$_\alpha$' and Dean says the corresponding thing about Nixon but with a possibly different subscript, then at least one of them fails to include the other's remark within the scope of his own. The situation with iterations of the determinacy operator is quite different: e.g. if Nixon says 'Everything Dean says is D^α not true',[48] and Dean says the corresponding thing about Nixon, then both succeed in disagreeing with the other's remark even though they have used the same ordinal α. Because of this, there is little pressure to employ high ordinal superscripts on determinacy operators in normal contexts. And because of that, it is difficult to find circumstances where it is plausible to maintain that we should have reason to think that a person's theory is 'defective in some sense of that term' without there being a

[47] Or alternatively, as 'either *defective* or $\neg D\neg$*defective* or $\neg D^2\neg$*defective* or . . . ' , where '*defective*' means 'neither determinately true nor determinately false'. In some G-logics this is a slightly stronger predicate than the one in the text.

[48] I focus on the case of 'D^α not true' rather than 'not D^α true' to maintain the parallel with 'not true$_\alpha$' in the stratified truth theory. If $\alpha < \beta$, 'true$_\beta$' is weaker than 'true$_\alpha$' but 'D^β' is stronger than 'D^α'; so 'not true$_\alpha$' gets stronger as α increases, whereas 'not D^α' would get weaker. (However, the immediate claim in the text would hold just as well for 'not D^α true'.)

sufficiently high α—say $\in_0$, or the first non-recursive ordinal, or even higher—for which the operator D^α is perfectly clear and for which we are in a position to think the theory to be not D^α *True*.

Part Four Transcending the Hierarchies?

13 A new revenge worry, in three strengths

Part Four will in effect be concerned with the question (raised in the last section) of why we can't use the truth predicate to 'unify' the various iterations of the determinacy operator into a ' hyperdeterminately' operator D_{hyp}. This is closely related to a revenge worry: for if we *could* formulate such an operator, then we could formulate a 'Hyper-Liar' sentence (one which asserts its own lack of hyper-determinate truth), which couldn't be handled along the lines used for the various 'determinate Liar' sentences in the hierarchy $\{Q_\alpha\}$, and perhaps couldn't be consistently handled at all without giving up the truth and satisfaction schemas. Of course, from known consistency results, there can't be an operator D_{hyp} definable in the language whose corresponding Liar sentence forces a violation of the truth or satisfaction schemas; but it is prima facie puzzling why we can't use the truth predicate to create such an operator.

But the concern in Part Four will not be only with the idea of hyper-determinately operators that are *definable in the language*; the real worry is that the hierarchy of determinately operators used in working out a G-solution can be used to *make such a hyper-determinately operator intelligible*. We may not be able to define such an operator in the language, because of certain expressive limitations of the language; but the idea is that we can transcend these limitations in the mind ('mentally quantifying' over the levels of the hierarchy), and then contemplate *adding such an operator to the language*. The **strong revenge worry** is that adding such an operator to the language would produce a new paradox that requires giving up the truth schema. Substantiating this would be a fatal blow to any claim that a G-solution adequately resolves all the paradoxes.

There are weaker revenge worries also based on the idea that we can use the hierarchy of determinately operators to make intelligible some sort of hyper-determinately operator. The **intermediate-strength revenge worry** is that such an operator can't be made consistent with the truth schema *in the semantic framework so far introduced*, that is, in a G-semantics. If that were right it would mean that if we were to expand the language to include these new concepts, we would get new paradoxes *that can't be resolved by the same sort of means by which paradoxes were resolved in the language L that has been treated* (though they might be resolved by other means). That too would undermine any claim that G-solutions offer an ultimate answer to the paradoxes. The natural way to try to argue for the intermediate-strength revenge worry is to argue that we

can make intelligible an operator that is both idempotent and satisfies the inferential versions of the conditions on being a very weak determinacy operator (see Section 10); for we've already seen that such an operator couldn't possibly be treated within a G-semantics.

The **weak revenge worry**—so weak that maybe it shouldn't count as a revenge worry at all—is that we can use the hierarchy of determinately operators to make intelligible some sort of hyper-determinately operator not definable in the original language, which breeds new prima-facie paradoxes. It is *not* claimed that these new paradoxes aren't resolvable in a G-semantics—that would be the intermediate-strength worry. So it wouldn't really undermine the claim that we need nothing more than G-semantics to resolve the paradoxes. Still, if the weak worry could be substantiated it would show that we couldn't make do with a G-semantics for a single language: a G-solution for a single language would generate a richer language that needs its own G-solution, and so forth. That wouldn't defeat the basic idea of G-solutions, but it would be a disappointment. It would show that *one* of the disadvantages of classical stratified truth theories carries over to G-solutions. (G-solutions would however still retain the first and third advantages discussed in the previous section; indeed, it is arguable that even the second would not be *totally* undermined.[49])

As I've noted, the only obvious way to try to argue for the intermediate-strength worry is to argue for the intelligibility of an *idempotent* determinacy operator (or at least, an idempotent 'very weak determinacy operator'). For the weak revenge worry, this is not so: we could substantiate it by quantifying over all iterations of *D that are expressible in the language L*, and there's no obvious reason why the result of so doing should be idempotent. If it isn't idempotent, then in an expanded language L^* that includes it, we'd get a new hierarchy obtained by iterating D_{hyp}; that's how a G-semantics for L^* might be possible (and hence why a substantiation of the weak worry would only be a relatively minor blow to G-solutions).

The prima-facie case for a D_{hyp} operator being idempotent (and thus supporting at least the intermediate-strength worry) seems slight:

(i) We would get idempotence from the assumption that hyper-determinacy claims obey excluded middle (together with conditions (b_w) and (c_w) of Section 10). But it isn't evident how excluded middle for hyper-determinacy claims could be argued, short of either the assumption that excluded middle holds generally (which would of course rule out G-solutions from the start, independent of revenge worries)

[49] Part of the reason is implicit in the point made at the very end of the previous section: there is normally little practical need to ascend very high in the iterations of *D*, in strong contrast to the case of stratified classical theories.

or the assumption that we can read a hyper-determinacy claim off the model theory (an assumption that I hope to have disposed of in Section 9).

(ii) We could plausibly get idempotence from the assumption that a hyper-determinacy predicate would unify *all* iterations of *D*, *even those not expressible in the language L*. In particular, suppose that one of the 'iterations of *D*' included in the 'unification' would be 'hyper-determinately hyper-determinately'; then 'hyper-determinately' would have the full-strength of 'hyper-determinately hyper-determinately', and so (assuming condition (c_w) of Section 10) 'hyper-determinately' would be idempotent. But there seems little basis for the thought that we can define a hyper-determinateness operator that unifies *even those iterations of D not expressible in the language L*. (Perhaps we could get such an argument from the assumption that 'is an iteration of *D*' is a bivalent predicate; but we'll see that that assumption is wholly unwarranted.)

I might add that even if there were an idempotent hyper-determinacy operator, that wouldn't seem to support the *strong* revenge worry: presumably one could avoid paradox in various ways that are consistent with keeping the truth schema, e.g. by denying the iterability of the operator or by in some other way restricting the inference from *A* to $D_{hyp}A$. Admittedly, such solutions are unattractive; and my view is that the rationale for a G-solution would be thoroughly undermined if it could be argued that an idempotent determinacy operator is intelligible.

These last few paragraphs concern the intermediate and strong revenge worries; but as I've said, my goal is to argue that there is no basis for *even the weak form* of revenge worry. But a word of clarification is in order. It is certainly not part of my claim that it is incoherent to imagine that the language *L* be expanded in non-definitional ways. There is always a possibility of introducing new concepts; this is certainly the case for concepts pertaining to new forms of society or new organisms or newly discovered particles, and I see no reason to doubt that it is so for new mathematical concepts as well. And it may well be that adding new mathematical concepts to the language could make further iterations of the determinacy operator expressible in the language, and consequently would lead to an extension of the G-solution to the enriched language. (Extending the G-solution would be completely mechanical once one had the more powerful mathematical concepts.) I take it that *that* wouldn't count as any kind of problem for a solution to the paradoxes, and so I'm understanding the 'weak revenge worry' to require more than this. What it requires to substantiate the weak revenge worry is that *simply by reflecting on the hierarchy of determinacy operators* we can make intelligible an operator that transcends them; it is that, and not merely that we might make this intelligible in some other way, that seems to generate 'levels of language' in some objectionable sense. This is vague—the weak form of the revenge

worry *just is* vague.[50] But I think that what follows will undermine the worry, and in the process undermine the intermediate and strong worries too.

14 Hierarchies of operators: introductory remarks

In order to properly discuss this, we need to be much clearer about transfinite hierarchies of iterations of a determinacy operator.

Any iteration D^α of D will be a syntactic operator that takes any L-formula A to an L-formula $D^\alpha A$ with the same free variables.[51] For any formula A, we can take $D^0 A$ to just be A; that is, we can take D^0 to simply be the identity operator on L-formulas. If we've defined the operator D^α, then we can define the operator $D^{\alpha+1}$: for any formula A, the result of applying $D^{\alpha+1}$ to A is to be the result of applying D to the result of applying D^α to A.

So the only issue in specifying the hierarchy of iterations of D is specifying D^λ for limit λ. Of course, we want to do this in such a way that for any formula A, $D^\lambda A$ is in effect the infinite conjunction of the $D^\alpha A$ for $\alpha<\lambda$. There will be a limitation on how far we can do this, so let us say that we want it for all limit ordinals λ less than a certain limit ordinal σ; σ will then be called the length of the hierarchy.

More precisely, let OP be the set of operations on formulas of L that assign to each formula another formula with the same free variables; and for any $O \in OP$, let $det(O)$ be the operation that takes each formula x into the result of applying D to Ox.

Definition: A *hierarchy (of iterations of D)* is a function H from $\{\alpha \mid \alpha < \sigma_H\}$ to OP, for some limit ordinal σ_H, that meets the following conditions:

RCZ $H(0)$ is the identity operator;
RCS For any $\alpha<\sigma$, $H(\alpha + 1)$ is $det(H(\alpha))$.
RCL For any limit ordinal $\lambda<\sigma_H$, $H(\lambda)$ is a member of OP such that for any L-formula x and any assignment s of objects to the free variables of x, s satisfies $[H(\lambda)](x)$ if and only if for all $\beta < \lambda$, s satisfies $[H(\beta)](x)$.

[50] The intermediate and strong forms can be freed of similar vagueness, for they can be taken to involve intelligible expansions of the language however achieved.

[51] Each D^α is a recursive operator, since once its application to a specific formula such as '$0 = 0$' is specified, there is a mechanical procedure for turning that into a specification of its application to any other formula A. (The mechanical procedure is intuitively a kind of "generalized substitution": it involves not only substituting A for appropriate occurrences of '$0 = 0$', but also substituting the standard name of A for appropriate occurrences of the standard name of '$0 = 0$', the standard name of the standard name of A for appropriate occurrences of the standard name of the standard name of '$0 = 0$', and so on.) Since each such operator is recursive, it is definable in the 'true'-free fragment L_0 of L. (The word 'true' will occur in sentences that are *mentioned* in the definitions of D^λ for limit λ, but that is just syntax and doesn't prevent the definition being in L_0.) Of course, the fact that each specific D^α is definable in L_0 doesn't show that the whole hierarchy of them is; and the hierarchies I will focus on will turn out to be far too large to be definable in L_0.

I will call these the *Reasonability Conditions* for the behavior of a hierarchy on zero, successors, and limits respectively. The condition (LIM) of Section 11 was just the model-theoretic analog of (RCL).

A minor complication here is that RCL uses 'satisfies' in a way slightly differently from the way used so far: it speaks of satisfaction of formulas *with arbitrarily many free variables* by *assignments of objects to the free variables*, whereas I have taken satisfaction to be of formulas *with a single free variable* by *objects*. There is nothing deep here: by a slight extension of the ideas in note 4, the use of 'satisfies' employed in (RCL) could be defined from my official one. (I spare you the details.)

Of course, 'satisfies' in this modified sense is still a non-classical notion: excluded middle cannot be assumed to hold generally for it. This gives the condition (RCL) a rather different character than conditions (RCZ) and (RCS). And it raises a point which will turn out to have major significance: **we have no obvious reason to think that 'is a hierarchy' is a predicate that obeys excluded middle.** Indeed, we'll see that the supposition that it obeys excluded middle leads to contradictions. However, we'll also see that we can define more restrictive notions of hierarchy for which excluded middle can be assumed.

Although this definition allows for a multiplicity of hierarchies (not only hierarchies of different lengths, but also different ones of the same length), it is *roughly* the case that different hierarchies of the same length are 'equivalent', and that those generated by paths of different lengths are ' compatible'. More precisely, call two operators O_1 and O_2 on L-formulas *equivalent* if for every L-formula A and every assignment function s, s satisfies O_1A iff it satisfies O_2A. Call two hierarchies H_1 and H_2 *compatible* if for every ordinal α in the domain of both, $H_1(\alpha)$ is equivalent to $H_2(\alpha)$. (And call two hierarchies *equivalent* if they are compatible and have the same length.) Then we have the following:

EquivalenceTheorem: H_1 and H_2 are hierarchies $\vDash$ H_1 is compatible with H_2.

Note that I have not stated this as a conditional, and can't do so in general because of the problem with excluded middle. (Recall that in G-logics, $\rightarrow$-introduction is only valid on the assumption of excluded middle for the premise.) The equivalence claim in the first sentence of this paragraph did state it as a conditional, which is why I emphasized that it was only a rough approximation to the truth.

Proof of Equivalence Theorem: an obvious induction, but it is worth spelling out to ensure that no fallacious use is made of excluded middle. So, assume that H_1 and H_2 are hierarchies, A is any formula, and s is any assignment function. We need that for all $\alpha < \min\{\sigma_{H_1}, \sigma_{H_2}\}$, s satisfies $[H_1(\alpha)]A$ if and only if it satisfies $[H_2(\alpha)]A$. By the transfinite induction rule, which is valid even for non-classical predicates (see end of

Section 5), it suffices to show that for any $\alpha < \min\{\sigma_{H_1}, \sigma_{H_2}\}$,

> $(\forall \beta < \alpha)$(s satisfies $[H_1(\beta)]A$ if and only if s satisfies $[H_2(\beta)]A) \rightarrow$
> s satisfies $[H_1(\alpha)]A$ if and only if s satisfies $[H_2(\alpha)]A$.

Given the supposition that H_1 and H_2 are hierarchies, the conclusion is evident when α is 0 or a successor. For limit λ, the supposition that they are hierarchies gives that for any $\alpha < \min\{\sigma_{H_1}, \sigma_{H_2}\}$,

> [s satisfies $[H_1(\alpha)]A$ if and only if $(\forall \beta < \alpha)$(s satisfies $[H_1(\beta)]A)]\wedge$
> [s satisfies $[H_2(\alpha)]A$ if and only if $(\forall \beta < \alpha)$(s satisfies $[H_2(\beta)]A)]$,

which yields the desired conclusion via the inference $(X_1 \leftrightarrow Y_1) \wedge (X_2 \leftrightarrow Y_2) \vDash (Y_1 \leftrightarrow Y_2) \rightarrow (X_1 \leftrightarrow X_2)$, which is easily seen to hold in all G-logics by several applications of Condition (II) from Section 5. ⊠

A crucial question is, how far can we get hierarchies to extend? It's easy enough to satisfy the instances of (RCL) when λ is sufficiently simple: e.g. when λ is ω. When A is a formula with a single free variable 'v', we could let $D^{\omega}A$ be the formula

> For all n, v satisfies every formula that results by prefixing $\langle A\rangle$
> with n occurrences of D.

This has the same free variable that A does. In addition to its occurring freely, a formula that contains it is mentioned, but this is unproblematic. (And we can remove the restriction to formulas with only 'v' free: e.g. let $D^{\omega}A$ be

> For all n, every formula that results by prefixing $\langle A\rangle$ with n
> occurrences of D is satisfied by every
> assignment s of v_{i_1} to $\langle v_{i_1}\rangle$ and ... and v_{i_k} to $\langle v_{i_k}\rangle$;

where these are the free variables of A.)

So specifying D^{ω} is easy; but as the limit ordinal λ becomes more complicated, specifying D^{λ} becomes more difficult. Indeed it can't be done at all when λ is sufficiently large, for instance, when λ isn't definable in L or when any of its predecessors isn't definable in L.

Moreover, when it *can* be done, there will often be significantly different ways to do it, corresponding to different ways that λ and its predecessors might be defined. In particular, we might define λ as the supremum of a certain sequence S of ordinals; and then we might be able to take $D^{\lambda}A$ as saying roughly that for all ordinals α in S, the result of prefixing an appropriate operator D^{α} to A is true (or true of v, if A contains 'v' free). But if there is one sequence with λ as its supremum there will be many, so the precise choice of the operator D^{λ} will not be unique. And the non-uniqueness increases as λ increases, because then many of the D^{α} from which D^{λ} is defined will themselves not be unique. In short, a specification of D^{λ} that takes the form suggested

will depend on a whole path p of definitions of limit ordinals. Indeed, we'll see that it is possible to define a function Φ that, roughly speaking, 'takes paths to hierarchies': given that p is a path of definitions of limit ordinals, $\Phi(p)$ will be a hierarchy. I'll sometimes call it D_p, to emphasize that it is a path-dependent hierarchy of iterations of D. (I'll also use the notation $D_p{}^\alpha$ as a suggestive abbreviation for $[\Phi(p)](\alpha)$.) (I will give a more rigorous treatment of paths and path-dependent hierarchies in ensuing sections.)

There are many paths of a given length if there are any at all, and the hierarchies generated by distinct paths are distinct. It would be possible to live with this high degree of non-uniqueness of the hierarchies, given the equivalence theorem. But it is more convenient to restore as much uniqueness as we can. An obvious idea for doing so is to existentially quantify over the paths, i.e. to let $D^\alpha A$ be defined as something like '$\exists p[Path(p) \wedge \langle D_p{}^\alpha A\rangle$ is true]'. We'll see that as long as we put certain bounds on the length of the paths, or equivalently on the ordinal α, it is possible to fully restore uniqueness by such a route. But when we lift those bounds, uniqueness can't be fully restored without making the definition virtually worthless, and this is crucial to the dissolution of revenge problems.

15 'Small' hierarchies

I'll eventually want to consider hierarchies of iterations of D that extend 'as far as possible' through the ordinals.[52] But the fact that the notion of a hierarchy cannot be assumed to obey excluded middle will complicate the discussion, and so in this section and the next I will give a 'warm-up' that restricts to 'small' hierarchies in which this problem does not arise. More specifically, let L_R be some fixed fragment of L for which we know excluded middle to hold. (It might for instance be the 'true'-free fragment L_0, or the fragment L_1 consisting of sentences in which 'true' occurs only in the context 'true and a sentence of L_0'.) Let λ_R be the first ordinal that is not definable in L_R. The path-dependent hierarchies to be discussed early in this section can have lengths up to and including λ_R, and by the end of the section I will have unified them into a single path-independent hierarchy of length λ_R. Even when the fragment of L under consideration is L_0, the hierarchies to be considered are thus very much larger than those considered in [3]: in that paper I imposed very stringent requirements on the hierarchies, which entailed that their length had to be a recursive ordinal. I can now see no good motivation for those stringent requirements.

Let σ be any limit ordinal no greater than λ_R, i.e. any limit ordinal all of whose predecessors are L_R-definable. In particular, every *limit* ordinal less than σ is

[52] In the same sense that someone might want to be 'as rich as possible', even if he didn't think a state of 'maximal possible richness' makes sense.

L_R-definable. (The latter claim is really no weaker than the former, since there is an obvious way of obtaining a definition of a successor ordinal from a definition of the largest limit ordinal that precedes it.) So there is a function p—many of them in fact—that assigns to each limit ordinal λ less than σ some L_R-definition of it, i.e. some L_R-formula (with exactly one free variable) that is satisfied by λ and by nothing else. Call any such function p an *L_R-path of length σ*. Note that σ is determined by p: given any L_R-path p, its 'length' (in my slightly nonstandard sense) must be the first limit ordinal for which p is undefined, which I'll call σ_p. Also, note that 'L_R-path' involves the notion of L_R-definability and hence is not a term of L_R. But it is a term of the fragment L_{R^*} that results from L_R by allowing 'true' to occur in the context 'true and a sentence of L_R'; and excluded middle must hold throughout this fragment given that it does throughout L_R. So for any function p, either it is an L_R-path or it isn't.

To repeat, for any limit ordinal σ up to and including λ_R, there are L_R-paths p whose length is σ; and obviously this is not so for any $\sigma > \lambda_R$.

I now state a special case of a theorem proved in an Appendix to the paper. Let σ be any countable limit ordinal, let $Pred(\sigma)$ be the set of its predecessors, and $Pred_{\text{lim}}(\sigma)$ be the set of limit ordinals that precede it. Let P_σ be the set of functions from $Pred_{\text{lim}}(\sigma)$ to formulas with a single free variable 'μ' (which we can think of as a variable restricted to countable limit ordinals), and let P be the union of the P_σ for all countable σ. (P is thus an L_0-definable set to which we expect all L_R-paths to belong, whatever the fragment L_R.) Let OP be the set of operations on L-formulas, and let J be the set of functions from initial segments of the countable ordinals to OP. (J is thus an L_0-definable set to which we expect all hierarchies to belong.) Then:

> **Theorem on existence of small hierarchies:** There is an L_0-definable function $\Phi : P \longrightarrow J$ such that for every L_R-path p, $\Phi(p)$ (*aka* D_p) is a hierarchy of length σ_p.

Since we have just seen that there are L_R-paths of any limit length up to and including λ_R, we have:

> **Corollary:** There are path-dependent hierarchies of any length up to and including λ_R.

Let an *L_R-hierarchy* be a hierarchy of form $\Phi(p)$ for p an L_R-path. An obvious but important fact about the notion of an L_R-hierarchy (for a specific L_R within which excluded middle holds) is that it obeys excluded middle: any function either is an L_R-hierarchy or isn't one. The theory of L_R-hierarchies can thus by developed without attending to the subtleties caused by possible failures of excluded middle. Note that an L_R-hierarchy needn't be *definable* in L_R. Indeed, it needn't be definable even in the full language L, and it won't be if p itself is undefinable in L. One issue of some interest (though it won't be a central concern here) is: for which σ are there *L_R-definable*

hierarchies of length σ. It is immediate that for a hierarchy (or a path) to be L_R-definable, its length σ must be *strictly less than* λ_R: for λ_R is by definition undefinable in L_R, but if a hierarchy (or a path) is L_R-definable then so is its length. It can easily be shown that there is no maximal length for L_R-definable hierarchies, indeed, each L_R-definable hierarchy has an L_R-definable proper extension.[53] I don't know if there are L_R-definable hierarchies on arbitrarily large proper initial segments of $Pred(\lambda_R)$, but even in the case of L_0 we have L_0-definable hierarchies extending well into the non-recursive ordinals.[54]

Even though specific L_R-hierarchies needn't be definable in the full L, we can nonetheless quantify over all of them (including the undefinable ones) in the 'successor fragment' L_{R^*}. For sake of simplicity, I will define a hierarchy of operators that apply only to 'v'-formulas (formulas whose sole free variable is 'v'); this can be extended to a hierarchy of operators on arbitrary formulas by the route illustrated for D^ω in the previous section.

> **Definition of path-independent hierarchy of length λ_R** : If A is any 'v'-formula, let $D^\alpha_{[L_R]}A$ be the formula (whose free variables are 'α' and 'v')
>
> $\exists p$(p is an L_R-path $\wedge\ \alpha < \sigma_p\ \wedge$ the result of applying $[\Phi(p)](\alpha)$ to $\langle A\rangle$ is true of v).

(We could restrict the quantification to paths of length λ_R without affecting the result.) Since there are no L_R-paths of length greater than λ_R, $D^\alpha_{[L_R]}A$ is false of everything if $\alpha \geq \lambda_R$. Consequently, $D_{[L_R]}$ has useful application only when $\alpha < \lambda_R$; after this, it fails to meet condition (RCL) on being a hierarchy. But within this domain of useful application, $D_{[L_R]}$ behaves very nicely: for (i) using the Equivalence Theorem in the strong conditional form (which is legitimate in contexts like this where we have excluded middle for the antecedent), we have that for each L_R-path p, the operator

[53] Let H be an L_R-definable hierarchy. If its length is a successor $\gamma + 1$, extend it by adding $\langle \gamma + 1, det(H(\gamma))\rangle$. If its length is a limit ordinal λ (which must be L_R-definable, since H is), extend it by adding $\langle \lambda, O^*_H\rangle$, where O^*_H is the operator that assigns to each formula x whose sole free variable is 'v' the following formula:

> what results from substituting the L_R-definition of H and the standard name of x into the blanks in 'For all operators O in the range of __, v satisfies the result of applying O to __)'.

The definition of O^*_H really needs to be generalized to apply to arbitrary formulas, but one way to do that was illustrated in the discussion of D^ω in the previous section and another will be mentioned in the Appendix.

[54] For instance, let p be the function that assigns to the smallest non-recursive ordinal ν the formula 'μ is the smallest non-recursive ordinal', and assigns to each recursive ordinal the formula $j(\alpha)^\wedge$'is a Church-Kleene notation for μ', where $j(\alpha)$ is the *numerically smallest* Church-Kleene notation for it (in **O** or some other universal system). Then p is L_0-definable and is an L_0-path, with domain $\{\alpha \mid \alpha < \nu + 1\}$. From this we could easily get L_0-paths extending much farther, e.g. to the limit of ν, ν^ν, ν^{ν^ν}, ... (and beyond).

$D^{\alpha}_{[L_R]}$ is equivalent to D^{α}_{p} wherever the latter is defined; and (ii) for each $\alpha < \lambda_R$ there are L_R-paths for which it is defined. So we have a single quite natural hierarchy that usefully extends all the way up to λ_R.

16 A length-independent hierarchy? Revenge?

So far I've been holding the 'effectively classical' fragment L_R fixed, and seeing what can be done within it. But as I've noted, whenever we have a fragment L_R that we know to be 'effectively classical' , we can enlarge it to a 'successor fragment' L_{R^*}. The path-independent hierarchy $D_{[L_{R^*}]}$ usefully extends further than $D_{[L_R]}$ does: it extends up to λ_{R^*}, which is strictly larger than λ_R. In their common domain of usefulness, i.e. when $\alpha < \lambda_R$, the operators $D^{\alpha}_{[L_R]}$ and $D^{\alpha}_{[L_{R^*}]}$ will not be identical, for the formulas $D^{\alpha}_{[L_{R^*}]}A$ quantify over more paths. But the operators are nonetheless equivalent (when $\alpha < \lambda_R$): for any formula A, $D^{\alpha}_{[L_R]}A$ and $D^{\alpha}_{[L_{R^*}]}A$ are true of exactly the same things. In other words, the hierarchy $\{D^{\alpha}_{[L_{R^*}]} \mid \alpha < \lambda_{R^*}\}$ is *in effect* a proper extension of the hierarchy $\{D^{\alpha}_{[L_R]} \mid \alpha < \lambda_R\}$ (though it doesn't *literally* extend it since it assigns different formulas at each infinite stage).

In short, though we have achieved a certain kind of path-independence, we have not achieved length-independence: given any path-independent hierarchy of the sort described in this section, we can convert it to a longer one. We have a 'hierarchy of path-independent hierarchies' .

But isn't there a way to produce a unique hierarchy by 'unifying' the ones we have? One might be tempted to argue as follows:

> Consider the set S of all λ_R for classical fragments L_R; there is a smallest limit ordinal ρ such that $\rho \geq \lambda_R$ for all λ_R in S. So for each $\alpha < \rho$, there are $\lambda_R \in S$ for which $\alpha < \lambda_R$; pick one, and let D^{α} be $D^{\alpha}_{[L_R]}$. This defines a hierarchy extending up to ρ which is guaranteed to be well-behaved (since at each stage it is equivalent to a well-behaved operator).

But this argument presupposes that there is such a set as S, which in turn presupposes that there is such a set as 'the set of all (effectively) classical fragments L_R' . But that supposition is justified only if we can assume excluded middle for the predicate 'is a classical fragment'. And the assumption of excluded middle here is both prima facie unwarranted and demonstrably inconsistent.

It is prima facie unwarranted because to call a fragment effectively classical is to say that for each formula A within it, excluded middle holds. But we know from the end of Section 3 that we can't assume excluded middle for claims of form 'A obeys excluded middle': indeed, from excluded middle for 'A obeys excluded middle' one can infer excluded middle for A (note 14). If we can't assume excluded middle for all claims A, why should we be able to assume it for 'L_R is a fragment all members of which obey excluded middle'?

But the key point is that it is actually *inconsistent* to assume excluded middle for the predicate 'is a classical fragment'. The reason for that is that *given that additional assumption* the argument displayed above is valid, and easily leads to the further conclusion that the 'unified hierarchy' of length ρ is itself effectively classical. But then we can extend the hierarchy past ρ, by applying the Small Hierarchy-Existence Theorem to a path of length ρ. So it would follow that there is an effectively classical hierarchy bigger than all effectively classical hierarchies and so bigger than itself.

The fact that 'effectively classical fragment' is not a predicate for which excluded middle can be assumed makes it difficult to find useful generalizations of the form 'For all effectively classical fragments . . . ' . That is why in this section I have avoided doing so. Rather, I began the section by noting that L_0 is clearly an effectively classical fragment, as are L_1, L_2, and so forth; indeed, for each clear case L_R of such a fragment, its 'successor fragment' L_{R^*} is one too. That is enough to give the neverending 'hierarchy of path-independent hierarchies' that I have discussed.

Consequences for revenge? Suppose that we pick a particular path-independent hierarchy in the 'hierarchy of hierarchies' ; say $\{D_{[L_R]}{}^{\alpha} \mid \alpha < \lambda_R\}$. Are we faced with even a weak form of revenge problem? Obviously not: we have defined $D_{[L_R]}{}^{\alpha}$ *for variable* α, so we can easily 'unify' the operators $D_{[L_R]}{}^{\alpha}$ in *this* hierarchy simply by defining $D_{hyp(L_R)}A$ as

$$\forall\alpha(\alpha < \lambda_R \rightarrow D_{[L_R]}{}^{\alpha}A).$$

$D_{hyp(L_R)}$ is not a member of the hierarchy $\{D_{[L_R]}{}^{\alpha} \mid \alpha < \lambda_R\}$, but it is definable in L (and indeed, in L_{R^*}), by the definition just given; indeed, it is just the $\lambda_R{}^{th}$ member of the proper 'extension' of that hierarchy $\{D_{[L_{R^*}]}{}^{\alpha} \mid \alpha < \lambda_{R^*}\}$. Since $D_{hyp(L_R)}$ is definable in L, the general consistency result applies to it: so we have a guarantee that $D_{hyp(L_R)}$ does not lead to paradox.

The point of this discussion is simply to serve as a warm-up for the discussion that follows of what can be done in the full L. Obviously the discussion so far does nothing to dispel the worries of Section 13: it simply shows that it is possible within the language to transcend the hierarchy of iterations D^{α} *for those* α *that are definable in a single demonstrably classical fragment of the language*. The results might even encourage the weak form of revenge worry: for the discussion shows that we can intelligibly transcend a hierarchy of iterations of $D_{[L_R]}{}^{\alpha}$ for α definable in L_R, but only by going to a higher language L_{R^*}; which might suggest that we can intelligibly transcend a hierarchy of all those iterations D^{α} of D that are definable in the full L, but only by going to a richer language. That is the issue to which I now turn.

17 General hierarchies

What happens when we go from iterations D^{α} of D definable in a single classical fragment of L to iterations definable in the full L? A crucial point will be that the

question of which syntactic operators on sentences count as iterations of *D* becomes 'fuzzy': or put more precisely, we cannot in general assume

$$(O \text{ is an iteration of } D) \vee \neg(O \text{ is an iteration of } D).$$

To see why this is so, let's define 'α^{th} iteration of *D*' as best we can for α larger than those discussed in the previous section.

To this end, we generalize the notion of path introduced early in Section 15. (Recall the definitions given there of *P* and *J*.) Let an *L-path* be some member *p* of *P* that assigns to each limit λ less than σ_p some *L*-definition of it. We know from Section 4 that the concept of an *L*-definition is 'fuzzy', i.e. we can't in general assume excluded middle for claims of form '*u* is an *L*-definition of *v*'; so there is no evident reason to assume it for formulas of form '*p* is an *L*-path'. (Any *L*-path *p* is an ordinary function on $Pred_{\text{lim}}(\sigma)$ for some unique σ, and we can reason about such functions in normal ways; but the question of which such functions count as *L*-paths is 'fuzzy'.) We can also generalize the Hierarchy-Existence Theorem:

General theorem on existence of hierarchies: There is an L_0-definable[55] function $\Phi : P \longrightarrow J$ such that

p is an *L*-path $\vDash \Phi(p)$ (*aka* D_p) is a hierarchy of length σ_p.

(Again, the proof is deferred to the Appendix.)

But note that this theorem is a much weaker result than we had for the more restricted sorts of paths discussed in the previous section: it *doesn't* say that (for every *p*) *if p* is an *L*-path *then* D_p is a hierarchy extending up to σ_p. The function Φ 'constructs' *something* from every $p \in P$, but because it may be 'fuzzy' whether *p* is an *L*-path, it also may be 'fuzzy' whether what's been constructed is a hierarchy of iterations of *D*. We can't even say that *if p is an L-path then* what we've constructed is such a hierarchy; all we can say is that if we're in a position to assert that *p* is an *L*-path then we're in a position to assert that what we've constructed is such a hierarchy.

We do have:

Corollary 1: $\exists p$(*p* is an *L*-path of length σ) $\vDash \exists H$(*H* is a path-dependent hierarchy of iterations of *D* with length σ).

Proof: The previous theorem gives

$$p \text{ is an } L\text{-path of length } \sigma \vDash D_p \text{ is a hierarchy of } D\text{-iterations extending up to } \sigma.$$

Existentially generalize over D_p in the conclusion, then use $\exists$-introduction on *p*. ⊠

[55] The fact that Φ is definable in a classical fragment of *L* is of little intrinsic interest, but is essential to the proof: the proof employs an inductive definition that relies on the Replacement Schema, which is suspect once we leave demonstrably classical fragments of *L*.

But as with the previous theorem, we can't infer that *if* all predecessors of σ are L-definable *then* there is a hierarchy of D-iterations extending up to σ.

In addition, there is less of a connection than we might expect between the premise of the corollary and the claim that all predecessors of σ are L-definable. Certainly if there is an L-path of length σ, then all predecessors of σ are L-definable; but for the converse we are restricted to

Lemma: All predecessors of σ are L-definable $\models$ There is an L-path of length σ.

Proof: The premise implies that $(\forall \lambda \in \mathit{Pred}_{\lim}(\sigma))(\exists y)(y$ is an L-definition of $\lambda)$; so by the 'choice principle' mentioned at the end of Section 5 (and valid even for predicates not assumed classical), there is a function p with domain $\mathit{Pred}_{\lim}(\sigma)$ such that $(\forall \lambda \in \mathit{Pred}_{\lim}(\sigma))(p(\lambda)$ is an L-definition of $\lambda)$. ⊠

Call an ordinal **almost hereditarily L-definable** if all its predecessors are definable in L. (Note that ordinals that are (fully) hereditarily definable count as *almost* hereditarily definable as well.) Then

Corollary 2:

(Negative Part) If σ is not almost hereditarily L-definable then there are no path-dependent hierarchies of iterations of D with length σ.

(Positive Part) σ is almost hereditarily L-definable $\models$ there are path-dependent hierarchies of iterations of D with length σ.

Proof: The Positive Part comes from the Lemma and Corollary 1, and the negative part is immediate. ⊠

We can also define a path-independent 'hierarchy' ${D_{|L|}}^{\alpha}$ in complete analogy to how we defined the various ${D_{|L_R|}}^{\alpha}$; we'll see, though, that it is much less tractable. (Again I restrict the definition to 'v'-formulas, for simplicity.)

Definition of general path-independent 'Hierarchy': If A is any 'v'-formula, let ${D_{|L|}}^{\alpha} A$ be the formula (whose free variables are 'α' and 'v')

$\exists p(p$ is an L-path $\wedge\ \alpha < \sigma_p \wedge$ the result of applying $[\Phi(p)](\alpha)$ to $\langle A\rangle$ is true of $v)$.

This defines the 'hierarchy' for arbitrary α, but as with the hierarchies ${D_{|L_R|}}^{\alpha}$ of the previous section, there comes a point when it becomes ill-behaved: indeed, it eventually becomes trivial, in that for every sufficiently large α, ${D_{|L|}}^{\alpha} A$ is false of everything, for every formula A. What makes the situation much worse in this case is that we can say very little about where the breakdown occurs; indeed, this will turn out to be a 'fuzzy' question.

We do have the following:

General path-independent hierarchy theorem:

Negative part: If λ is a limit ordinal that is not definable in L, then for any $\alpha \geq \lambda$, ${D_{|L|}}^{\alpha}$ is trivial. Consequently, if σ is not almost hereditarily L-definable, then $D_{|L|} \restriction \mathit{Pred}(\sigma)$ fails

(very badly!) to be a genuine hierarchy of iterations of D (or to be a genuine hierarchy of reasonable candidates for determinacy operators).

Positive part: σ is almost hereditarily L-definable $\vDash D_{[L]} \restriction Pred(\sigma)$ is a genuine hierarchy of iterations of D.

(The proof of both parts is almost immediate from Corollary 2.) Note that since there are countable ordinals with undefinable predecessors, the negative part of this theorem implies that $D_{[L]}$ becomes very badly behaved *in the countable ordinals*. And the positive part, being in rule form rather than conditional form, is not enough to allow us to conclude that $D_{[L]}$ satisfies (RCL) up to any ordinal all predecessors of which are L-definable. For in order to universally generalize, you need a conditional to universally generalize on, and the theorem above does not license the strengthening to conditional form. The best we have is this: **for each limit σ that *we are in a position to assume* has only L-definable predecessors, we can take $D_{[L]} \restriction Pred(\sigma)$ to be a genuine hierarchy of iterations of D.** Once you *prove* that a given limit σ has only L-definable predecessors, the above results *converts the proof* into a demonstration that $D_{[L]} \restriction Pred(\sigma)$ is an adequate hierarchy.

18 Maximal hierarchies?

A question of great interest to revenge worries is whether there is a *maximal* hierarchy of iterations of D, that is, a σ for which $D_{[L]} \restriction Pred(\sigma)$ is adequate as a hierarchy (i.e. satisfies (RCL)) but $D_{[L]} \restriction Pred(\sigma + \omega)$ isn't. I will give a proof of the following 'negative' answer:

Anti-maximality theorem: The assumption of such a maximal hierarchy of iterations of D is inconsistent.

In the course of this I will also establish the less interesting claim

Lemma for anti-maximality theorem: The assumption of a maximal *L-definable* hierarchy of iterations of D is inconsistent.

I take the Lemma to have little intrinsic interest to the revenge problem: the proponent of revenge is sure to argue that the concepts that give rise to revenge problems aren't definable in the language. But the Anti-Maximality Theorem goes against not merely the assumption of a maximal *L-definable* hierarchy of iterations of D: the definability of the hierarchy doesn't enter into the result.

Proof of Lemma: If $D_{[L]} \restriction Pred(\sigma)$ is a genuine hierarchy of iterations of D then all predecessors of σ are definable in L (by negative part of Path-Independent Hierarchy Theorem). Assume that the hierarchy $D_{[L]} \restriction Pred(\sigma)$ is definable. Then σ is definable, as the length of the hierarchy. But then $\sigma + n$ is definable too for all finite n, and so

every ordinal less than $\sigma + \omega$ is L-definable. But then the positive part of the Path-Independent Hierarchy Theorem tells us that $D_{[L]} \restriction Pred(\sigma + \omega)$ is adequate as a hierarchy of iterations of D. And this hierarchy is definable in L: for $\sigma + \omega$ is definable since σ is, and we've defined $D_{[L]}$. So $D_{[L]} \restriction Pred(\sigma)$ isn't maximal among the definable hierarchies of D-iterations. And so the assumption that it is maximal among the definable such hierarchies, and the existential generalization of that assumption, are inconsistent. ⊠

Proof of theorem: Assume that $D_{[L]} \restriction Pred(\sigma)$ is a maximal hierarchy of iterations of D. Then we can define σ as the largest limit ordinal λ for which $D_{[L]} \restriction Pred(\lambda)$ is a hierarchy of iterations of D. But then we can define $D_{[L]} \restriction Pred(\sigma)$, so it is a maximal definable hierarchy, which is inconsistent by the Lemma. ⊠

It is important to be clear that from the inconsistency of the claim that there is a maximal hierarchy of iterations of D, it doesn't follow that there is no such hierarchy. (Any more than it follows from the inconsistency of the truth of the Liar sentence that the Liar sentence isn't true.) And indeed, the negation of the maximality claim is inconsistent too.[56] So the maximality claim has a status very much like that of the Liar sentence, in that the assumption that there *either is or isn't* a maximal hierarchy of iterations is inconsistent. We are in the realm of the 'inherently fuzzy'. (Anyone tempted to think that there *must* be a way to 'unify' the hierarchies into a maximal one should re-read the response in Section 16 to the argument that there must be a way to 'unify' the effectively classical hierarchies.)

19 Hyper-determinacy and revenge

In Section 13 I distinguished three strengths of revenge worry. The weak form, which would not be totally devastating if substantiated, was that once one had a hierarchy of determinacy operators in a language L, one would be naturally led to a new 'hyper-determinacy' operator, not definable in L but intuitively meaningful; since it is in an expansion L^* of L, the consistency proof for L wouldn't directly apply (though it might be extended to L^* in a completely mechanical way). The more serious revenge worries were the 'intermediate strength' worry that a consistency proof for L^* would have to be quite different from that of L, and the 'strong' worry that consistency for L^* could only be achieved by giving up the truth or satisfaction schema. To substantiate these stronger worries, the intelligibility of a hyper-determinacy operator would not suffice: we would need that operator to be idempotent (and even given that, there would be no obvious case for more than the intermediate worry). We are now in a

[56] Since each hierarchy has only L-definable ordinals in its domain, the lack of a maximal hierarchy would imply that there are arbitrarily large initial segments of the ordinals all members of which are definable in L, and that is absurd on cardinality grounds.

position to see that even the weak form of the worry was unfounded: there is no way to generate an understanding of a notion of hyper-determinacy from the hierarchy of determinacy notions that we have in the language.

We've seen that we can define within L a 'hierarchy' $D_{[L]}$ that can be extended as far as one likes; but this is not a hierarchy of iterations of D—nor a hierarchy of operators that are in any intuitive sense determinacy operators—since it eventually starts mapping every sentence into a falsehood. (This breakdown occurs somewhere in the countable ordinals.) Moreover, it is inconsistent to assume that there is a maximal initial segment of the ordinals on which $D_{[L]}$ behaves adequately—that is, a maximal fragment $D_{[L]}{\restriction}\mathit{Pred}(\sigma)$ such that for every limit ordinal $\lambda < \sigma$, ${D_{[L]}}^{\lambda}$ is equivalent to the 'conjunction' of all the ${D_{[L]}}^{\beta}$ for $\beta < \lambda$.

Since this is inconsistent, it seems that the best we can do if we want to avoid the danger of choosing a 'hierarchy' that is inadequate is to choose a hierarchy that we can show to be adequate. But this will always be less than maximal. Given such a less than maximal hierarchy $\{{D_{[L]}}^{\alpha} \mid \alpha < \sigma\}$, we can always quantify over the operators in its range: calling a sentence A ' hyperdeterminately true' would then be saying

(H_σ) For all α less than σ, $\langle {D_{[L]}}^{\alpha} A\rangle$ is true.[57]

(σ is bound to be definable in L if the fragment is adequate.) But in formulating (H_σ) we are in effect just going another step in a longer hierarchy *that is already in the language*. The general consistency proof for the theory of truth in L applies to everything expressible in the language, including iterations of D longer than those in the specific non-maximal hierarchy $\{{D_{[L]}}^{\alpha} \mid \alpha < \sigma\}$. So it certainly applies to sentences containing 'hyper-determinacy' claims if that simply means claims of form (H_σ), for those are in the language

There is however another possibility to consider: to put it picturesquely, we can introduce the idea of a 'fuzzy initial segment' of the full 'hierarchy' $D_{[L]}$; in particular, the 'initial segment consisting of all and only those ordinals that are in some adequate hierarchy', where again an adequate hierarchy (what I earlier called a genuine hierarchy) is one that obeys (RCL). This picturesque way of speaking makes no literal sense: since it is inconsistent to assume excluded middle for 'adequate', talk of an 'initial segment' defined via the notion of adequacy is simply ill-defined. Even so, there is an idea behind the picturesque talk that can be made intelligible: that we define a 'hyper-determinacy' operator using a quantifier restricted by the predicate 'is an adequate hierarchy'. I distinguish two versions of this:

$[H_{\supset}]A$ $\quad \forall\alpha(\alpha$is in the domain of an adequate hierarchy $\supset \langle {D_{[L]}}^{\alpha} A\rangle$is true);
$[H_{\rightarrow}]A$ $\quad \forall\alpha(\alpha$is in the domain of an adequate hierarchy $\rightarrow \langle {D_{[L]}}^{\alpha} A\rangle$is true).

[57] This should really be written as 'For all α less than σ, $\langle {D_{[L]}}^{\nu} A\rangle$ is true of α', but I trust the more readable notation in the text will not confuse.

It should be clear from the start that explaining hyper-determinateness in either of these ways can't possibly serve the purposes of the person who wants to argue for a revenge problem, even a weak one: for both $H_{\supset}$ and $H_{\rightarrow}$ are already in the language L. Since they are in the language, they can't possibly lead to paradox, given the general consistency proof. But it is worth seeing how the arguments for paradox fail. In the case of both operators, the failure of the paradoxical arguments points up the fact that 'fuzzily restricted quantifiers' have to be treated with extreme care.

We saw in Section 10 that no operator E in the language can jointly satisfy four conditions. Those conditions were expressed in terms of models as (b_w), (c_w), (d), and idempotence; the corresponding inferential conditions are

(c_w) $\models EA \rightarrow A$
(d) $A \rightarrow \neg A \models \neg EA$
(b_w) $A, \neg EA \models \bot$
(Idem) $\models EA \rightarrow EEA$ (or equivalently, $\models \neg EEA \rightarrow \neg EA$)

The reason the conditions aren't jointly satisfiable (to transcribe an argument from Section 10 into inferential terms) is that for any such operator E we can formulate a sentence Q_E that asserts its own lack of E-truth, so that $\models Q_E \leftrightarrow \neg EQ_E$. Then since (d) implies $EQ_E \rightarrow \neg EQ_E \models \neg EEQ_E$ we'd have $EQ_E \rightarrow Q_E \models \neg EEQ_E$; so using ($c_w$), $\models \neg EEQ_E$. Idempotence would then yield $\models \neg EQ_E$, hence $\models Q_E$. (b_w) would then yield $\models \bot$, which is impossible. Given this general result, we know that neither $H_{\supset}$ nor $H_{\rightarrow}$ can possibly satisfy all of these four conditions. The question is, which of these do they satisfy, and which of the desirable additional conditions (a), (b), and (c) do they satisfy? (The D^{α} operators in adequate hierarchies satisfy the full (b) and (c) in addition to (a) and (d). I take that to be highly desirable: (b_w) and (c_w) were singled out only as minimal conditions for inconsistency with (Idem) and (d); and while the retreat from (c) to (c_w) is perhaps within the bounds of acceptability, the retreat from (b) to (b_w) would be a major one.)[58] I won't investigate (c), but will say enough about the other principles to show that neither $H_{\supset}$ nor $H_{\rightarrow}$ are operators that have much appeal.

(In discussing these matters I will occasionally state things in terms of standard set-theoretic models for L. It's worth noting that in the definitions of $H_{\supset}$ and $H_{\rightarrow}$

[58] Incidentally, we might want to add a further condition: that the operator E is to strengthen a given determinacy operator D, i.e.

(c_w*) $\models EA \rightarrow DA$.

Given this, E is bound to satisfy both (c_w) and (d), since D does. And we can now show that (c_w*) and (b_w) are together incompatible not only with (Idem) but with the weakened form of it

(W-Idem) $\models EA \rightarrow DEA$.

The proof is an obvious modification of the one given: (c_w*) yields $\models \neg DEQ_E$, which with (W-Idem) yields $\models Q_E$, which with (b_w) yields absurdity.

we could take all quantification over ordinals to be restricted to countable ordinals; for this reason, the points raised in Sections 8-9 about the 'misleadingness of models' in dealing with sentences with unrestricted quantifiers will not affect the ability to infer truth from having value 1 and falsity from having value 0; conversely we can take *clear* truths to have value 1 in these models and *clear* falsehoods to have value 0.)

Let's start with $H_{\supset}$. Here the principles (c_w) and (d) are valid (in the extended sense introduced at the end of Section 5). The reason is that the *L*-definability of 1 is valid, as is the equivalence of $True(\langle D^{1}_{[L]}A\rangle)$ with DA; so the conditional $H_{\supset}A \rightarrow DA$ is valid. Since D satisfies (c_w) and (d), it is then evident that $H_{\supset}$ does so as well. (It can also be shown that $H_{\supset}$ satisfies condition (a).)

An inspection of the above proof shows that because $H_{\supset}$ satisfies the principles (c_w) and (d), the sentence $H_{\supset}H_{\supset}Q_{\supset}$ must be false (where $Q_{\supset}$ is the 'Liar sentence' corresponding to $H_{\supset}$). (Model-theoretically, $H_{\supset}H_{\supset}Q_{\supset}$ must have value 0 in any standard set-theoretic model.) $H_{\supset}Q_{\supset}$, on the other hand, will not have value 0 in any such model: for it is equivalent to $\neg Q_{\supset}$, and $Q_{\supset}$ won't have value 1. The reason: $Q_{\supset}$ says

$\neg\forall\alpha$(α is in the domain of an adequate hierarchy $\supset \langle D_{[L]}{}^{\alpha}Q_{\supset}\rangle$ is true);

that is,

$\exists\alpha$(α is in the domain of an adequate hierarchy $\wedge\neg(\langle D_{[L]}{}^{\alpha}Q_{\supset}\rangle$ is true)),

and the only way for this to have value 1 in a model that satisfies the truth schema is for it to be the case that for some α, $|\alpha$ is in the domain of an adequate hierarchy$| = 1$ and $\left|D_{[L]}{}^{\alpha}Q_{\supset}\right| = 0$. But $|\alpha$ is in the domain of an adequate hierarchy$| = 1$ only if α is in the domain of a clearly adequate hierarchy, in which case we can't have $|Q_{\supset}| = 1$ and $\left|D_{[L]}{}^{\alpha}Q_{\supset}\right| = 0$. The contradiction shows that $H_{\supset}Q_{\supset}$ can't be 0; so idempotence fails and paradox has been blocked. (By the reasoning of note 58, even 'Weak-Idempotence' fails: that is, $DH_{\supset}$ is strictly stronger than $H_{\supset}$.)

So if the operators $D_{[L]}{}^{\alpha}$ are deemed problematic because not idempotent, $H_{\supset}$ offers no advantage. But in fact $H_{\supset}$ is far worse than the operators $D_{[L]}{}^{\alpha}$ in an adequate hierarchy, for it violates condition (b) in an extreme way: **For any sentence** A **whatever,** $H_{\supset}A$ **is 'at best fuzzy'.** (In any reasonable model for the language, no sentence of form $H_{\supset}A$ gets value 1).[59] The reason is that $H_{\supset}A$ amounts to the claim '$\forall\alpha$($\neg$(α is *L*-definable) $\vee$ $D_{[L]}{}^{\alpha}A$ is true)'; but this claim can't be assumed true even when A is a clear truth like '$0 = 0$', for when 'it is fuzzy whether α is *L*-definable' it will be 'fuzzy whether $D_{[L]}{}^{\alpha}A$ is true', so that the disjunction '$\neg$(α is *L*-definable) $\vee$ $D_{[L]}{}^{\alpha}A$ is true' will itself be ' fuzzy'. In short, $H_{\supset}A$ will never be clearly true.

[59] Of course, given that $H_{\supset}$ (i) is already in a language that has a G-semantics and (ii) is not syntactically restricted in its application, it must be value-functional; and this means that if there is any sentence A with value 1 for which $|H_{\supset}A| < 1$, this *must* be so for every sentence with value 1.

We see, then, that though $H_{\supset}$ is a well-defined operator, it is quite a worthless one; it does not correspond to a notion of determinacy in any reasonable sense.

What about $H_{\rightarrow}$? Here the situation seems to be even worse. I say 'seems to be' because sentences of form $H_{\rightarrow}A$ are extremely complicated—this becomes evident when one consults the Appendix to see how the path-dependent hierarchies used in defining $D_{[L]}$ are themselves defined—and I have not been able to come up with a rigorous argument settling exactly how $H_{\rightarrow}$ behaves. However, I think there is *very* strong reason to believe the following:

1. In many G-logics, including all the published ones and all possible ones that satisfy the very natural law (III_s) of note 23, $H_{\rightarrow}$ is a completely trivial operator, in that $H_{\rightarrow}A$ is clearly false for every sentence A. (So in a standard set-theoretic model of such a logic, $|H_{\rightarrow}A|$ is always 0, even if A has value 1.)
2. In any other G-logic, $H_{\rightarrow}$ will share the problem of $H_{\supset}$: $|H_{\rightarrow}A|$ will never have value 1, creating a serious disadvantage as compared with the genuine hierarchies of form $D^{\alpha}_{[L]}$. In these logics, $H_{\rightarrow}$ will also fail to be idempotent, thus destroying the entire rationale for going beyond hierarchies of form $D^{\alpha}_{[L]}$. ($H_{\rightarrow}$ as defined may also fail to satisfy (c_w), though that problem could be fixed by conjoining the above definiendum for $H_{\rightarrow}A$ with A.)

To substantiate 2, it would suffice to show that in any standard set-theoretic model for L, there are ordinals α for which $|\alpha$ is in the domain of an adequate hierarchy$| \not\leq |\exists p(p$ is a path of length greater than $\alpha \wedge \langle D^{\alpha}_{p}A\rangle$ is true)|.[60] To substantiate 1, it would suffice to show that in the semantics for the logics mentioned in 1 we can strengthen this, replacing '$\not\leq$' by '$>$'. And I think a strong case can be made for these claims, though I will not attempt it here since the case depends on a careful look at the way in which the $D^{\alpha}_{p}A$ are defined in the Appendix. (The fact underlying the plausibility of the claims is that $|D^{\alpha}_{p}A|$ is evaluated by looking at $|D^{\beta}_{p}A|$ for all β that precede some ordinal μ for which $\left|\mu \text{ satisfies } \_\_\_\right| > 0$, where the blank is filled by the 'attempted definition of α' that p assigns to α. For ordinals α for which $|\alpha$ is in the domain of an adequate hierarchy$|$ is between 0 and 1 and hence the 'attempted definition' is not clearly adequate, we can expect this to include ordinals β much bigger than α, perhaps some for which $|\exists p(p$ is a path of length greater than $\beta)|$ is 0 and hence $|D^{\beta}_{p}A|$ will be 0.)

I've claimed only a *strong case* for the claims 1 and 2. But a strong case isn't a proof; what if I'm wrong? If I'm wrong, that would have some interesting ramifications: it would mean that the range of iterations of D available in the language is far richer than Section 17 might have suggested. More particularly, we could continue the sequence of non-idempotent operators even further than was done in the general hierarchies considered there. This would not however change anything substantive

[60] In analogy with n. 57, I use $|F(\alpha)|$ as a more readable notation for $|F(v)|_{\alpha}$.

in what I've said: by the results of Section 11, the new hierarchy would never lead to idempotence as long as we could iterate in accordance with (RCL); and in accordance with the result of Section 18, there would be no maximal hierarchy of iterations of $H_{\rightarrow}$. And so introducing the operator $H_{\rightarrow}$ wouldn't have served the purpose that an advocate of 'revenge' or of a 'unified determinacy operator' intended. (It might motivate such a person to introduce a new *hyper-hyper* determinacy operator; but obviously nothing of philosophical significance could be gained by further travel down this road.)

The crucial point is that even if my conjecture that $H_{\rightarrow}$ fails to meet the conditions for being a determinacy operator is wrong, this couldn't possibly generate even a weak revenge problem, because $H_{\rightarrow}$ is already in the language. $H_{\rightarrow}$ avoids the most obvious threat of paradox by failing to be idempotent. And the general consistency result shows that it can't possibly lead to any paradox, because it is in *L*.

20 Conclusion

The discussion of the last few sections gives strong reason to think that we simply have no conception of any genuine hyper-determinacy operator that isn't definable in *L*: the closest we can come is operators like $H_{\supset}$ and $H_{\rightarrow}$ that *are* definable in *L* but that don't behave like a hyper-determinacy operator (or any sort of determinacy operator) ought to behave.

I haven't tried to argue that there is no intelligible expansion of our understanding of a hierarchy of determinacy operators. Indeed, it is clear from my formal constructions that if we were to expand our mathematical language in such a way that countable ordinals not hereditarily definable in our current *L* (or even, not *clearly* hereditarily definable) were to become clearly hereditarily definable, then that would expand our conception of the hierarchy: it would enable us to make stronger 'iterated determinacy' claims than we can make today. But such an expansion of mathematics couldn't be simply a matter of defining new concepts in terms of current vocabulary, it would have to involve coming to have concepts that can't be clearly defined in the existing language; and achieving such an expansion is no simple task.[61]

In particular, part of the upshot of my argument is that such an expansion can't be achieved simply by reflecting on the hierarchies of determinacy operators already definable in the language. The thought that reflecting on such hierarchies leads to a concept of hyper-determinacy that transcends the language is simply an illusion, an illusion created by the failure to appreciate that if we try to quantify over all the

[61] One might wonder about imaginary beings with an *uncountable* language that contained a name for every countable ordinal. The results in this chapter could be extended to them: it's simply that the hierarchies would extend into the uncountable ordinals.

determinacy operators definable in the language, the range of the quantification is indeterminate.

This, I think, defeats the (slightly vague) weak revenge worry of Section 13. *A fortiori* (and much more important), it defeats the idea that reflection on such hierarchies leads to a concept of an *idempotent* hyper-determinacy operator of the sort that might support an intermediate-strength (or perhaps even strong) revenge worry.

Of course, one might think that there is a case for thinking that we can understand an idempotent determinacy argument that is independent of reflecting on the hierarchies we can define. I've mentioned four possible grounds for this thought, three in Section 13 and another in Section 12, and I'll now add a fifth.

- One is the model-theoretic revenge argument, which I believe I have refuted in Section 9.
- Another is the thought that excluded middle holds generally. This is certainly not a view I can claim to have refuted in this chapter: all I have tried to do (in Section 2) is to sketch the costs that the semantic paradoxes raise for such a view, and to elaborate a view that seems on balance to have less drastic costs. (Of course, one may evaluate costs differently: *de gustibus non disputandum*.)
- A third one is that excluded middle should at least hold for claims of the form 'O is a reasonable candidate for being a determinately operator'. If that were true, we should be able to unproblematically quantify over all reasonable candidates for determinately operators, to produce a hyper-determinately operator that obeys excluded middle; it is then a short step to idempotence. But why think excluded middle holds for claims of this sort? The discussion in the last few sections provides strong reasons to doubt this supposition, and it's hard to imagine a case for the supposition that doesn't rely either on excluded middle generally or on the thought that the only reasonable candidate for a determinacy operator could be read off the model theory.
- A fourth argument (the one not directly mentioned before) is that even if excluded middle doesn't hold for claims of the form 'O is a reasonable candidate for being a determinately operator', still we should be able to quantify over all reasonable candidates for determinately operators to produce a hyper-determinately operator; it may not obey excluded middle, but it might be thought to be idempotent on other grounds. But from the results of the preceding section, I think we can reasonably extrapolate to the view that there is little reason to expect such an operator to be idempotent, and little reason even to think that it will obey minimal conditions for being a determinacy operator.

The final argument (the one from Section 12) is perhaps the one with most intuitive force: it is that we *just need* a unified notion of determinacy or defectiveness. Note however that this argument cannot very well be advocated by the classical theorist,

since the classical theorist has no such unified notion either. Nor can it very well be advocated by the proponent of any other solution to the paradoxes in which such a notion is unavailable. Indeed, I'm not sure that there are any demonstratively consistent theories (or even non-trivial dialetheic ones) that have such a notion available and hence are in a position to advocate this argument. I'm willing to concede (for the moment anyway) that it would be a point in favor of a solution to the paradoxes that it had a unified notion of defectiveness. If there are ways to achieve this that don't have overwhelming costs, they should be developed and weighed against the solutions to the paradoxes sketched here.[62]

Appendix: Proof of the Hierarchy-Existence Theorems

It's simplest to prove directly the existence of hierarchies of operators on formulas with a single free variable '*v*' ('*v*'-formulas). The Reasonability Conditions in the definition of the hierarchy are then modified in the obvious way: we restrict to operations on '*v*'-formulas, and in (RCL) we speak of satisfaction by objects instead of by assignment functions. From such a 'restrictive hierarchy' of operators on '*v*'-formulas, a more general hierarchy of operators on all *L*-formulas could be obtained: for by a minor extension of the ideas of note 4 we could define in L_0 a function that takes operations on '*v*'-formulas into corresponding operations on all *L*-formulas. (I spare you the details.)

The Hierarchy Existence Theorems of Sections 15 and 17 (modified in this way to apply to hierarchies of operators on '*v*'-formulas) follow almost directly from a technical lemma. Take P to be as defined in Section 15: recall that it is an L_0-definable set each member p of which is a function with domain the set of limit ordinals that precede σ_p, for some limit ordinal σ_p that may depend on p.

Hierarchy-construction lemma: There is an L_0-definable function $\mathbb{W}$ with domain $\{\langle p, \alpha\rangle \mid p \in P \wedge \alpha < \sigma_p\}$ that satisfies the following conditions:

1. For any $p \in P$, $\mathbb{W}(p, 0)$ is the identity operator on '*v*'-formulas.
2. For any $p \in P$ and any $\alpha < \sigma_p$, $\mathbb{W}(p, \alpha + 1)$ is $det(\mathbb{W}(p, \alpha))$.

[62] A number of people have tried to persuade me over the last few years that the 'revenge-immune' account in [3] doesn't really evade revenge. I should especially mention Graham Priest, who has mostly pressed model-theoretic revenge (see [17]) and Kevin Scharp, who has pressed the 'incompatibility qualm' and the 'counterintuitiveness qualm' of Section 12 (see section A.5 of [20]). It was in thinking about what Scharp says that I was led to realize that the discussion of the hierarchy of determinacy operators in [3] was insufficiently inclusive, and needed to be extended into the 'fuzzily definable' ordinals. I'm also very grateful to Josh Schechter for carefully reading an earlier draft and providing valuable comments that led me to substantial improvements.

3. For any $p \in P$ and any limit ordinal $\lambda < \sigma_p$, $W(p, \lambda)$ is an operator on 'v'-formulas such that for any 'v'-formula x, $[W(p, \lambda)](x)$ is equivalent to the 'v'-formula $Z_{p,\lambda}(x)$ that results from substituting $p(\lambda)$ and the standard name of x into the blanks in

$\forall\beta[\exists\mu(\_\_\_ \wedge \beta < \mu) \rightarrow$ the result of applying $W(p, \beta)$ to $\_\_\_$ is true of v].

Writing $W(p, \alpha)$ as $[\Phi(p)](\alpha)$, the Lemma implies that for every $p \in P$, $\Phi(p)$ satisfies the reasonability conditions for zero and successors (again, modified to apply to operators on 'v'-formulas). For the limit condition, on the other hand, we get the horrible-looking

For any $p \in P$ and any limit ordinal $\lambda < \sigma_p$, $[\Phi(p)](\lambda)$ is an operator on 'v'-formulas such that for any 'v'-formula x, $[[\Phi(p)](\lambda)](x)$ is equivalent to the 'v'-formula $Z_{p,\lambda}(x)$ that results from substituting $p(\lambda)$ and the standard name of x into the blanks in

$\forall\beta[\exists\mu(\_\_\_ \wedge \beta < \mu) \rightarrow$ the result of applying $[\Phi(p)](\beta)$ to $\_\_\_$ is true of v].

However, from the assumption that p is an L-path, what results from filling in the first blank is true of λ and nothing else, so $Z_{p,\lambda}(x)$ is satisfied by just those objects that satisfy all of the results of applying $[\Phi(p)](\beta)$ to x, for each $\beta < \lambda$. In other words, $\Phi(p)$ satisfies (RCL). Similarly for L_R-paths; but since we have excluded middle for 'L_R-path' (unlike for 'L-path), we can in the case of L_R convert the proof to the proof of the conditional: if p is an L_R-path then $\Phi(p)$ satisfies (RCL).

It remains only to prove the technical lemma.

Proof of hierarchy-existence lemma: The obvious idea for proving the Lemma is to define the function W using transfinite recursion. But a *direct* recursive definition of W seems impossible, because of the fact that condition (3) of the Lemma doesn't just *use* W (as in a normal recursive definition) but *mentions* it (by referring to a formula that contains it). So instead, I will recursively define a more generalized function F (with little intuitive meaning, I regret to say), then use a fixed point argument to get the desired W.

Let Y be the set of formulas of L that have only the two variables 'β', and 'z' free. We want to recursively define a function $F(p, \alpha, e)$ for $p \in P$, $e \in Y$ and $\alpha < \sigma_p$, whose values are operators on 'v'-formulas. The idea is that if we then instantiate on an appropriate instance e_0, then the formula $F(p, \alpha, e_0)$ will serve as the desired W (and so for any specific L_0-path p, $F(p, \alpha, e_0)$ will serve as the desired hierarchy).

The recursive definition:

- For any $p \in P$ and $e \in Y$, let $F(p, 0, e)$ be the identity operator on 'v'-formulas.
- For any $p \in P$ and $e \in Y$ and any $\alpha < \sigma_p$, let $F(p, \alpha + 1, e)$ be $det(F(p, \alpha, e))$.
- For any $p \in P$ and $e \in Y$ and any limit ordinal $\lambda < \sigma_p$, let $F(p, \lambda, e)$ be the operator that assigns to each 'v'-formula x the result of substituting $p(\lambda)$, e, and the standard name of x in that order into the blanks in the following schema:

  $\forall z\forall\beta[\exists\mu(\_\_\_ \wedge \beta < \mu) \wedge z$ is a syntactic operator on 'v'-formulas $\wedge \_\_\_ \rightarrow$ the result of applying z to $\_\_\_$ is true of v].

It should be noted that recursive definition is not unrestrictedly valid in L: it depends on the Replacement Schema, which is valid only in the context of excluded middle. But there is no problem with this particular recursive definition, for it is given in the 'true'-free fragment L_0. (The expression 'true' does occur here, but only in a sentence that is mentioned rather than used; it's mere syntax.) The recursive definition can be converted to an explicit definition of a relation $F(p, \alpha, e) = z$. Obviously for any particular e_0 that we restrict to, the first two bulleted conditions of the Lemma will be satisfied (by virtue of the corresponding conditions of the inductive definition); the task is to choose an e_0 that will make the third condition satisfied as well.

To this end, we now employ the Gödel–Tarski fixed point theorem on the formula '$F(p, \beta, e) = z$', to get a function $W(p, \beta)$ (defined in L_0) for which

$$\forall p \forall \beta [W(p, \beta) = F(p, \beta, \langle W(p, \beta) = z \rangle)].^{63}$$

Using '$W(p, \beta) = z$' to instantiating the e in above recursive definition, the limit condition of the definition yields

- For any $p \in P$ and any limit ordinal $\lambda < \sigma$, $W(p, \lambda)$ is an operator that assigns to each 'v'-formula x the result of substituting $p(\lambda)$, the definition of '$W(p, \beta) = z$', and the standard name of x in that order into the blanks in the following schema:

  $\forall z \forall \beta [\exists \mu (\_\_\_ \wedge \beta < \mu) \wedge z$ is a syntactic operator on L-formulas $\wedge \_\_\_ \rightarrow$ the result of applying z to $\_\_\_$ is true of $v]$.

So for any p, $[W(p, \lambda)](x)$ is equivalent to the result of substituting $p(\lambda)$ and the standard name of x into the blank in

$$\forall \beta [\exists \mu (\_\_\_ \wedge \beta < \mu) \rightarrow \text{ the result of applying } W_\sigma(p, \beta) \text{ to } \_\_\_ \text{ is true of } v],$$

which is Condition (3).

References

[1] Brady, Ross T. (1989). 'The non-triviality of dialectical set theory'. In Graham Priest, Richard Routley, and Jean Norman (eds.), *Paraconsistent Logic: Essays on the Inconsistent*, pp. 437–70. Philosophia Verlag.

[63] The most familiar form of the fixed point theorem applies to formulas. Applying it to the formula '$F(h, \beta, e) = z$', we get a three-place formula $G(h, \beta, z)$ of L_0 such that

$$\forall z [G(h, \beta, z) \leftrightarrow F(h, \beta, \langle G(h, \beta, z) \rangle) = z].$$

But $G(h, \beta, z)$ defines a function; writing it as $W(h, \beta) = z$, we get the claim in the text.

[2] Feferman, Solomon (1984). 'Toward useful type-free theories', I. *Journal of Symbolic Logic* 49: 75–111

[3] Field, Hartry (2003). 'A revenge-immune solution to the semantic paradoxes'. *Journal of Philosophical Logic* 32: 139–77

[4] ——— (2003). 'The semantic paradoxes and the paradoxes of vagueness'. In JC Beall, (ed.), *Liars and Heaps*, pp. 262–311. Oxford University Press

[5] ——— (2004). 'The consistency of the naive theory of properties' *Philosophical Quarterly* 54: 78–104

[6] ——— (2005). 'Variations on a theme by Yablo'. In JC Beall and Brad Armour-Garb (eds.), *Deflationism and Paradox*, pp. 53–74. Oxford University Press

[7] ——— (2006). 'Compositional principles versus schematic reasoning'. *The Monist* 89: 9–27

[8] ——— (2006). 'Maudlin's "truth and paradox" '. *Philosophy and Phenomenological Research* 73: 713–20

[9] Friedman, Harvey and Sheard, Michael (1987). 'An axiomatic aproach to self-referential truth'. *Annals of Pure and Applied Logic* 33: 1–21

[10] Gupta, Anil and Belnap, Nuel (1993). The *Revision Theory of Truth*. MIT Press, Cambridge, Mass.

[11] Kripke, Saul (1975). 'Outline of a theory of truth'. *Journal of Philosophy* 72: 690–716

[12] Leeds, Stephen (1978). 'Theories of reference and truth'. *Erkenntnis* 13: 111–29

[13] Maudlin, Tim (2004). *Truth and Paradox*. Oxford University Press, Oxford

[14] McGee, Vann (1991). *Truth, Vagueness, and Paradox*. Hackett, Indianapolis

[15] Montague, Richard (1963). 'Syntactic treatments of modality, with corollaries on reflexion principles and finite axiomatizability'. *Acta Philosophica Fennica* 16: 153–67

[16] Priest, Graham (1987). *In Contradiction*. Martinus Nijhoff, Dordrecht

[17] ——— (2005). 'Spiking the field artillery'. In JC Beall and Brad Armour-Garb (eds.), *Deflationism and Paradox*, pp. 41–52. Oxford University Press

[18] Quine, W. V. O. (1970). *Philosophy of Logic*. Prentice-Hall, Englewood Cliffs, NJ

[19] Reinhardt, William (1986). 'Some remarks on extending and interpreting theories with a partial predicate for truth'. *Journal of Philosophical Logic* 15: 219–51

[20] Scharp, Kevin (2005). *Truth and Aletheic Paradox*. diss., University of Pittsburgh

[21] Soames, Scott (1999). *Understanding Truth*. Oxford University Press, Oxford

[22] Welch, Philip (forthcoming). 'Ultimate truth v. stable truth'. *Journal of Philosophical Logic*

[23] Williamson, Timothy (1994). *Vagueness*. Routledge, London

[24] Yablo, Stephen (2003). 'New grounds for naïve truth theory'. In JC Beall (ed.), *Liars and Heaps*, pp. 312–30. Oxford University Press

5

Validity, Paradox, and the Ideal of Deductive Logic

Thomas Hofweber

5.1 Thinking About Plan B

There are three basic ingredients that together give rise to the semantic paradoxes, and correspondingly there are three basic strategies for a straightforward solution to them. The first ingredient is a logic, like classical logic. This we can simply take as a set of inference rules associated with certain logically special expressions. The second is a truth predicate, or something like it, and some rules or schemas that are associated with it, say introduction and elimination rules, or the Tarski schema. The third ingredient is some other expressive resources that allow one to formulate one of the trouble inducing sentences, like the liar sentence, or the Curry conditional. Once we have these three we can derive contradictions and anything whatsoever using only the rules we have, with the help of the problematic sentences.

The three main strategies for a straightforward solution see the problem either in the logic, in the rules governing the truth predicate, or in the problematic sentence. Maybe we have to weaken classical logic, or the rules governing the truth predicate, or maybe the problematic sentences are not of the kind that the rules properly apply to. Any such solution will hold that one or another of the rules that we naively hold to be valid really isn't valid, or that for some reason the rules don't apply to the problematic sentences. Such a solution to the paradoxes I will call *a plan A solution*. It would simply

solve them by showing where our reasoning that leads us into trouble went wrong. It would point out which inference rule is not valid after all, and thus which rule is not to be reasoned with.[1] Most, if not all, solutions to the liar paradox are plan A solutions. And there is a tremendous amount of progress in finding such a solution. We have supervaluational solutions, McGee (1991), paraconsistent solutions,[2] Priest (2006), contextualist solutions, Simmons (1993), Field's sophisticated new solutions, Field (2003) and Field (2006a), and way too many more to list. But there is a growing sense among some of us that none of this is going to work in the end as a solution to the philosophical problem that the paradoxes pose. Without question there is much to learn from all these solutions and much to admire in their sophistication, but will they solve the philosophical problem that the paradoxes present? Here the main source of doubt is the existence of revenge paradoxes. It appears that even the most sophisticated solutions face further paradoxes that can be formulated using the terms of the solution. The justification for why certain rules are to be rejected as invalid will use certain new semantic terminology that then can be used to state new semantic paradoxes, ones especially tailored to the solution of the old paradoxes in question. If this is indeed so then one is faced with a dilemma. Either one pushes the problem simply somewhere else, or one has to insist that the solution does not apply to a language like our natural languages. But either way, the philosophical problem that the paradoxes pose is then not solved, it is simply pushed around.

To be sure, whether the standard plan A solutions all face revenge paradoxes is controversial, and some prominent proponents of such solutions explicitly claim that their solutions are revenge immune. I don't want to argue here that we can always find another revenge paradox, or that plan A solutions are to be given up. I personally have my doubts about them, but that shouldn't count for much. I simply want to ask in this chapter what if this is right? What if plan A solutions won't solve the paradoxes? Are there other options? Is there a plan B?

There is certainly one other option available, which isn't much of a plan, so it doesn't count as plan B. Maybe the paradoxes have no solution and this shows that our conceptual schema of describing the world and our talk about it in terms of truth and falsity, or warranted and unwarranted belief, simply collapses. If the reasoning in Curry's paradox is simply correct then it seems that every statement is true, and every statement is false. And if correct inferences transmit warrant from the premises to the conclusion then it also shows that every belief is equally warranted. This we could call *the Great Collapse*, and

[1] I will ignore solutions that find the problems not in the rules but in the problematic sentences alone. This mostly won't be relevant to contrast plan A with plan B, and that's why we will leave it aside for the most part.

[2] Of course, such solutions accept that it is fine to derive a contradiction, the error comes once you try to derive anything whatsoever from that contradiction.

it would be the greatest imaginable disaster for the project of inquiry. But maybe there is another possibility. Maybe the Great Collapse can be avoided even if plan A solutions fail. Let's call *a plan B solution* to the paradoxes a solution that avoids the Great Collapse and is not a plan A solution. This would have to be a solution that does not hold that one of the inference rules, either a logical one or one governing the truth predicate, is to be rejected, nor that the problematic sentences aren't appropriately instantiated in these rules. Rather a plan B solution takes the rules to be valid and the problematic sentences to be well formed and meaningful, but avoids the Great Collapse nonetheless. This might seem clearly impossible since if all the rules of classical logic plus the introduction and elimination rules for the truth predicate are valid, and the problematic sentences are allowed, then anything follows. But this, I think, is a mistake and ignores one option to respond to the paradoxes that to my knowledge has been neglected.[3]

In this chapter I would like to outline how such a plan B solution can go. I believe that it does not face revenge issues that bring down plan A solutions, and that it is generally attractive. In fact, I have my money on such a solution. According to the plan B solution to be outlined below, the real culprit is our conception of deductive logic as aiming for a certain ideal which is a philosopher's dream, but one we can live without. According to this plan B solution the problem isn't that we have the wrong rules, either logical or for the truth predicate, but rather that we have a wrong conception of what it is to have a deductively valid rule. I will outline how this could go in this chapter, I say more about it in Hofweber (2007), and I hope to develop it in more detail in the future.

5.2 Deductive Logic, Default Reasoning, and Generics

The rules of classical logic[4] are rules of a deductive logic. That is to say that any inference that is licensed in this system is valid and thus truth preserving, and the same holds for the inference rules governing the truth predicate. In fact, being truth preserving can be seen as the defining feature of a valid inference rule in a deductive logic:

[3] The notions of a plan A and plan B solutions do not quite correspond to Schiffer's notions of a happy face and unhappy face solution to the paradoxes. A happy face solution is much like a plan A solution in that it aims to uncover the error in our reasoning that leads into trouble. But for Schiffer an unhappy face solution is one that accepts defeat and then proposes a revision of our concepts that doesn't lead us into trouble. The plan B solution to follow is not revisionist in this way, and thus I prefer a different terminology. See Schiffer (2003).

[4] I will stick to classical logic in the following, since I will hold that it can be defended in the face of paradox even together with the full Tarski schema for the truth predicate. Thus the paradoxes don't force us to give it up. There might be other reasons to favor a different logic, but everything I will say about classical logic below works for basically any other logic as well.

(1) Inference rules are valid iff they are truth preserving.

But then, how can there be a plan B solution? If the inference rules are truth preserving and we consider the instances of them that lead to Curry's paradox or the liar paradox we simply get the Great Collapse. How is there even conceptually room for a way out? To see that there is more to say here, let's make a slight digression into rules that are used in ordinary, everyday reasoning that are commonly contrasted with deductively valid rules.

Ordinary, everyday reasoning mostly is not strictly deductive. When I think about what I should do when I see a bear in the wild I will reason with information that I represent as

(2) Bears are dangerous.

To accept (2) is closely tied to accepting an inference rule. It is the rule that allows one to infer from

(3) t is a bear

to

(4) t is dangerous

This is perfectly good ordinary reasoning, but it is not strictly deductively valid. Not all bears are dangerous. Some old bear without teeth might not be dangerous, but this does not make it the case that (2) isn't true. (2) is a generic statement. It means something like

(5) In general, a bear is dangerous.

Generic statements allow for exceptions. That is, the truth of a generic statement like (2) is compatible with the falsity of the corresponding universally quantified statement, in this case

(6) All bears are dangerous.

Nonetheless, they are closely tied to good inference rules, but these inference rules also allow for exceptions. Such inferences are ones that one is entitled to make unless one has overriding information. So, if I know nothing about Freddie except that he is a bear and I know (2) then I am entitled to infer that Freddie is dangerous. But if I learn in addition that Freddie is old and lost all his teeth I am not entitled to make that inference any more. Such reasoning is thus non-monotonic. More information can make an otherwise appropriate inference inappropriate.

This is commonly called default reasoning. By default I am entitled to make a certain inference, although more information can take that entitlement away from me. And generic statements like (2) closely correspond to inference rules in default reasoning. To accept a generic statement as true is closely tied to regarding such an inference in default reasoning as a good inference. All this is so even though the generic statement is not without exceptions, since not absolutely every bear is dangerous, and correspondingly, inference rules in default reasoning won't always be truth preserving since they can lead from the true premise that Freddie is a bear to the false conclusion that Freddie dangerous, which he is not any more. It is still perfectly rational to conclude that Freddie is dangerous from the premise that he is a bear even thought the rule I rely on in making this inference allows for exceptions, and even though I realize very well that it does so, and that not absolutely every bear is dangerous. There are many subtle features about generics and default reasoning which we won't be able to discuss here, but there is a rather straightforward lesson for our topic here. It is the key to having a plan B solution to the paradoxes.[5]

Default reasoning is commonly contrasted with reasoning that is deductively valid. Deductively valid inference rules are truth preserving in all instances and are monotonic whereas inference rules in default reasoning do not preserve truth in all instances and are non-monotonic. This is the ideal of deductive logic that I think we have to abandon. Even in deductive logic not all inference rules are always truth preserving, although it is rational to reason in accordance with them. I will propose that default reasoning and deductive reasoning are thus alike in this respect, although deductive logic can be different in various other ways from standard cases of default reasoning like the one discussed above. This is the lesson to be drawn from the paradoxes, it is the outline of a plan B solution to the paradoxes, and it in fact is much less radical than it might seem.

What seems to be distinctive of inference rules in deductive logic is that they are truth preserving, and above we considered it to be a defining feature of valid inference rules that they are truth preserving:

(1) Inference rules are valid if they are truth preserving.

This is hard to deny, and we should not deny it, properly understood. In fact, (1) has two readings. One of them, which we could call the strict reading, requires that

[5] The connection of generics to default reasoning is well known and widely discussed. See the introductory essay in Carlson and Pelletier (1995) for a survey of a number of topics about the semantics of generics, and Pelletier and Asher (1997) for a discussion of the relationship that generics have to default reasoning. We don't need the subtle details discussed in these articles for our main point in this chapter, so I am being brief here.

each instance is truth preserving. But (1) also has a generic reading. The right hand side has a generic reading which is nicely brought out by using the plural. Inferences are truth preserving according to the generic reading just like bears are dangerous. That doesn't mean that each and every instance is truth preserving, just like the latter doesn't mean that each and every bear is dangerous. But nonetheless, (1) is literally true on this reading. So, since the right hand side of (1) has two readings we should make this explicit and consider each reading a defining feature of two senses of validity:

(7) a. Let's call an inference rule *strictly valid* iff each and every instance is truth preserving.

b. Let's call an inference rule *generically valid* iff instances are truth preserving (understood as a generic statement).

The *ideal of deductive logic* holds that inference rules in deductive logic are strictly valid, and that this is the distinctive feature of *deductive* logic. The *default conception of deductive logic*, in contrast, holds that inference rules in deductive logic are only generically valid, although they might form a special subclass of the generically valid rules. I will propose that we should abandon the ideal of deductive logic and embrace the default conception of deductive logic instead. If we do so, we can have a plan B solution to the paradoxes, but there are other reasons to do so as well.

Suppose you take classical logic, in a rules only natural deduction version for now, as well as unrestricted introduction and elimination rules for the truth predicate. The default conception of this logic will hold that these rules are all valid, in the generic sense. And it will hold that it is rational to reason according to these rules, unless you have overriding reasons to the contrary in a particular case. This is exactly what we need to have a plan B solution to the paradoxes. According to the default conception of logic and the truth rules, all of them are valid and thus to be accepted. But there are instances of these rules that are not truth preserving. Curry's paradox is one, the liar paradox is another. These are the exception cases to the generically valid rules. They are the equivalent of the old toothless bear. Thus we can accept the rules, but rationally reject particular instances of them, and thus avoid the Great Collapse even though we accept classical logic and the truth rules. What we do have to give up is the ideal of deductive logic, but this isn't giving up very much. Thus the default conception of deductive logic gives one the tools to have a plan B solution to the paradoxes, at least in rough outline. In the next section I would like to elaborate on various aspects of this main idea. I won't be able to defend it in detail in this chapter, but I would like to discuss some of the issues that need to be addressed if one wants to defend it.

5.3 Spelling Out the Main Idea

Taking the rules of inference in deductive logic to be generically valid, but not strictly valid, gives us the possibility of a plan B solution to the paradoxes. Here are some questions that need to be addressed to develop this further, and how I intend to address them.

5.3.1 How is this a solution?

The proposal outlined above does not solve the paradoxes in the way a plan A solution aimed to do it. Plan A solutions basically try to find the mistaken inference rules that we used in reasoning ourselves into trouble, and to propose a framework or theory that shows how they are in fact mistaken. A plan A solution says what the mistaken inference step was, and why we made it nonetheless. On the present proposal each step in the reasoning that leads to paradox is correct in the sense that it is based on a rule which is (generically) valid and which is appropriately used in reasoning. So, in this sense there is no mistake in the reasoning that leads to paradox. But nonetheless, the present account holds, the conclusions drawn with the particular cases of the (generically) valid inference rules are rationally not to be accepted. In this sense it is a solution. Nothing went wrong in the reasoning. Each step is correct, in the sense that it is based on a (generically) valid rule of reasoning, a rule we are entitled to reason with, but the conclusion is still rationally not to be accepted. Thus we can accept the reasoning that leads us into paradox, but not accept its conclusions, or the Great Collapse. This is all the solution that the paradoxes need.[6]

And this captures exactly the natural reaction to the paradoxes. Ordinary reasoners accept all of the steps that lead to the paradoxes, they accept the relevant sentences with which we reason as perfectly well formed and meaningful, but they don't accept the conclusion of the reasoning without rejecting any of the steps. Now, the usual reaction from the philosophical side is that this is irrational and that there must be a mistake in the reasoning somewhere, one we philosophers will uncover. The present proposal sees a lot of wisdom in the natural reaction. And the generic conception of validity and inference rules shows how it is not irrational at all. Each step is in accordance with a valid rule, one that it is rational to follow, but the conclusion is nonetheless rationally not accepted.

[6] This solution only applies to the semantic paradoxes like the liar and Curry's paradox. It does not simply carry over to other paradoxes like, say, the Sorites paradox, or the paradoxes of motion, and so on. There are connections to other paradoxes as well, but in general I think there is more philosophical work to be done in these other cases than in our present case. In this chapter we restrict ourselves to the semantic paradoxes.

5.3.2 What are the exceptions?

If the inference rules of deductive logic and for the truth predicate are not strictly valid, but only generically valid, then the question arises: which cases are the exceptions to the strict validity? Which cases are the ones such that it is rational not to accept the conclusion derived from premises one accepts? Here the answer is quite straightforward for a believer in the generic conception of deductive logic, although this answer must be quite unsatisfactory for those who prefer plan A solutions to the paradoxes and who like the ideal of deductive logic.

The exceptions to the generically valid rules are simply the instances that don't preserve truth. This is, of course, not a very informative answer for those interested in the cases, but it is all that has to be said at the general level of spelling out the generic conception. Compare this to inference rules in default reasoning, like inferring from that t is a bear to that t is dangerous. Which are the exceptions to this rule? All the cases where t is a bear that is not dangerous, i.e. all the cases where this rule is not truth preserving. That is the right answer, but it doesn't help with the individual cases. If you want to find out if this particular bear is one which is an exception then you have to find out if he is dangerous. And the same holds for instantiations of the inference rules in logic with a truth predicate.

But then, why is it rational for us to not accept the instances that lead to the liar paradox, or the instances that lead to Curry's paradox? We are entitled to not accept them because we can clearly see that these are cases that don't preserve truth. In the Curry's paradox case it might preserve truth by accident, if the consequent of the relevant conditional is true. But we can see that this inference would have worked just as well for anything else, and thus if it did preserve truth it is an accident. Thus we can see that we are not entitled to the conclusion that we drew. This is quite clearly what we do realize when we think about the argument that leads to Curry's paradox, and because of this it is rational to reject this particular argument, although we accept all the rules of inference that were used in it. And similarly for the liar paradox.

5.3.3 How about revenge?

Plan A solutions seem to be threatened by revenge paradoxes, and this in part motivates plan B. But does plan B also lead to revenge paradoxes? Is it, too, only a way to push the problem somewhere else? Suppose we explicitly add the vocabulary of the present plan B solution to the language and we aim to formulate more paradoxes. Won't this cause trouble just as much as it did in the plan A cases?

Here plan A and plan B really differ. Plan A solutions will lead to revenge paradoxes unless they lead to expressive incompleteness. If they are expressively complete then we can formulate new sentences that lead to contradictions, or arbitrary conclusions. And

this shows that the particular plan A solution won't work. But for our plan B solution this is different. This solution already accepts that there are instances of the (generically) valid inference rules that lead to contradictions or arbitrary conclusions. But these instances are the exceptions to the (generically) valid inference rules. Any revenge paradox can only lead to more of these cases. A new super liar using the notions of generic validity or default reasoning could at best lead to further cases of instances of the inference rules that are exceptions to their (generic) validity. Since we already grant that there are such instances all that the revenge liar could show is that there are even further cases of the already accepted phenomenon. But this does not threaten our plan B solution. The threat of revenge for plan A solutions and the fact that it disappears for plan B solutions which are based on generic validity are a real advantage for plan B. There are other reasons to prefer plan B as well, but this certainly is one of the main ones.

5.4 How Radical Is It?

I suspect that there is a feeling that abandoning the ideal of deductive logic in favor of the default conception of deductive logic is simply going too far. Many people have suggested that radical consequences are to be drawn from the paradoxes, and maybe this is just another far out proposal that we have to give something up we clearly should try to keep. I don't think this is correct. In fact, what I am proposing that we give up has to be given up anyway, even for those who hold on to plan A solutions. And I don't think we are giving up that much in the end. The generic conception of deductive logic is good enough for most of the things that the ideal of deductive logic was supposed to do for us.

5.4.1 Counterexamples to the inference rules

If the paradoxes show that the inference rules in deductive logic are not strictly valid, it wouldn't be the only thing that shows this. In recent years a number of people have proposed counterexamples to various deductive inference rules, understood as rules for inferences in natural language. Vann McGee and Bill Lycan, for example, have argued that modus ponens fails for ordinary English conditionals.[7] Their arguments don't rely on the paradoxes, and if they are right then either modus ponens is to be rejected as a rule of inference, or it is to be modified in such a way that it is still acceptable, or it is to be understood as a generically valid rule. The generic conception of deductive logic has no problems with such examples, as long as, in general, modus ponens is valid. That there are exceptions does not refute the rule.

[7] See McGee (1985), Lycan (1993), and Lycan (2001).

5.4.2 Truth preservation and plan A

If the generic conception of deductive logic is correct then it is perfectly rational to accept an inference rule as valid even though one realizes very well that it does not always preserve truth. This might seem radical. However, Hartry Field has recently made a very good case that traditional plan A solutions to the paradoxes in fact have to accept just that. Field (2006b) considered the question of why a certain intuitive argument for the consistency of arithmetic fails. The argument is simply that since all axioms of arithmetic are true and inference preserves truth, all consequences of the axioms are true, and since all of their consequences are true, the axioms are consistent. Field notes that this argument has to break down somewhere, and depending on what one says about how a truth predicate is added to the underlying logic, there will be different places where it breaks down. In particular, Field argues that standard ways of adding a truth predicate without allowing for the deduction of everything can't maintain that all of the axioms are true and all of the rules are truth preserving. That is, given a certain system consisting of an underlying logic and a truth predicate governed by either axioms or rules that together avoid triviality, one of the following two options will hold for it: either it is a consequence of that system that one of its own axioms isn't true or that one of its own rules isn't truth preserving, or at least one can't consistently maintain from within that system that all the axioms are true and all the rules are truth preserving.[8] What this suggests is that even for plan A solutions, which are the ones Field discusses, one can't coherently hold that all the inference rules that one accepts are strictly valid. A plan A solution will either imply that there are instances of the rules that are not truth preserving, or at least will determine that one won't be able to consistently maintain that they are truth preserving. This takes quite a bit of the wind out of the sails of the criticism of the generic conception of deductive logic. Everyone will have to accept that it is rational to accept rules while one is at the same time unable to hold that all of their instances preserve truth.

Field's conclusion from his observation about why the intuitive argument against Gödel's second incompleteness theorem breaks down is partly congenial with the present approach, and partly in conflict with it. Field concludes that one should not think that the notion of validity of inference rules can be defined in terms of truth preservation. Rather it should be seen as a primitive, and it should be more closely associated with which rules it is rational to follow in inference. But this isn't the right lesson to learn, it seems to me. We can take

(1) Inference rules are valid iff they are truth preserving.

[8] Field notes that for a more limited case than what he discusses this can be read off some of the results of Friedman and Sheard (1987).

to define validity in terms of truth preservation, as long as the right hand side is understood as a generic statement. And if we grant that the truth of this generic statement gives us a good inference rule in default reasoning then we can see why this definition of validity makes valid rules good ones to reason with. And this is how it should be. To understand validity in terms of preservation of truth clearly gets something right, and to tie this to good rules to follow in inference clearly gets something right, too. The generic conception of deductive logic has these results.

5.4.3 How to live without the ideal

The ideal of deductive logic as involving strictly valid inference rules is based on the thought that there are forms of reasoning where we can never go wrong, no matter what the instances. This we have to give up. But almost everywhere where we rely on this ideal we could live with the generic conception of deductive logic just as well. Certain generic statements are tied to inference rules in default reasoning, and it is rational to reason according to such rules, although there are exceptions to them. In addition, when we reason according to rules in default reasoning we are warranted to hold the conclusions we draw using them, although we cannot hope to achieve absolute certainty this way. We might have used them on one of the exception cases, and it might be that what we concluded isn't true. But absolute certainty is neither required for knowledge nor for much else. In addition, it is not clear absolute certainty can be achieved even if the ideal were correct. Since the rules tied to generics are epistemically very much like strictly valid rules there will really be little difference.

Many other theses that are commonly associated with deductive logic carry over to the generic conception. For example, one can still hold that having a certain inferential role is constitutive of the meaning of the logical constants. But the inferential role has to be understood as figuring in certain generically valid inferences, not strictly valid inferences.

In addition, it will still be possible to draw an interesting distinction between deductive logic and other cases of default reasoning. The exception cases in deductive logic might be of a different kind, and there might be a special reading of the generic statement that is the appropriate one when we say that the inference rules are truth preserving. On a fuller development of the story which I can only outline in this chapter I hope to have more to say about this. In particular, I hope to distinguish different readings that generics can have and isolate the one that is relevant for characterizing in what sense deductive logic involves inferences that are truth preserving.

5.5 Schemas and the Truth-value of the Liar Sentence

Suppose what I have said so far is more or less correct. Then we can accept classical logic and the introduction and elimination rules for the truth predicate. Thus we get the features that come with classical logic in this setting, including

$$(8)\ p \vee \neg p$$

$$(9)\ \mathit{True}('p') \vee \mathit{True}('\neg p')$$

And to accept classical logic is to accept these schemas. So, what if we substitute the liar sentence for 'p'? Since it is classical logic one or the other disjunct has to be true. Which one is it?

Above we focused on a rule-centered version of classical logic. This doesn't have to be so, of course. We could start out with axiom schemas instead. But in either case, the question arises how we should understand schemas, either when they are axioms, or when they are derived using the rules and schematic formulas. As far as I can tell the believer in the generic conception of deductive logic has two options here. Both are strictly speaking available, but one is more congenial to the overall view than the other. Both options accept

(10) Instances of a schema (which is either an axiom or derived) are true.

but do so in different readings. (10) has a strict and a generic reading as well. According to the strict reading each and every instance is true. According to the generic reading instances are in general true, but there may be exceptions. On a strict reading it might well be that fewer schemas can be derived with the (generic) rules since particular instances of these rules aiming to derive a schema might turn out to be exceptions to the generically valid rules. Let's see what would happen if we take (8) or (9) to be derived, and what we should say about the truth-value of the liar sentence on each reading of instances of a schema that can be derived.

Suppose we accept the strict reading of schemas, that is, we take each and every instance of the schema to be true. Then this holds for the liar sentences as an instance. Since we assume classical logic we can ask which one of the disjuncts is true. One of them has to be true, but whichever one it is, it will lead to a contradiction. But any argument towards this contradiction will use some rules which are only generically valid. And the instances of these rules with either the liar sentence or its negation, or the claim that the liar sentence is true, or the claim that its negation is true, will be exceptions. They will lead to contradictions. Thus on this option either the liar or its negation is true, but one can't rationally conclude which one it is. It will be a case of ignorance, although there is an answer to the question. This option I take it is consistent with the generic conception of deductive logic, but maybe not as congenial with it as the next one.

Suppose, on the other hand, we accept the generic reading of schemas, i.e. instances of the schemas are true, understood as a generic statement. Then we can accept the schemas (8) as well as (9) and hold that instances of them with the liar sentence are exceptions. If one takes this route then the liar sentence will be neither true nor false, although the schema (9) is such that its instances are true, and the claim that the liar is either true or false is one of its instances. Again, this is no contradiction since we are assuming a generic reading of the claim that the instances of the schema are true. Of course, that means that

$$(11)\ \neg True(\ulcorner\lambda\urcorner) \wedge \neg True(\ulcorner\neg\lambda\urcorner)$$

which leads to contradictions in our classical setting, but this, again, will simply involve another case of an exception to the generically valid inference rules. This option is available to us as well. It denies that we are bound to ignorance of the truth-value that the liar has, since it is neither true nor does it have a true negation. This does not give rise to a revenge liar problem any more than the liar is a problem. It simply gives rise to different exceptions to the generically valid inference rules. And there are other options as well. I will remain neutral which option should be chosen in this chapter.[9]

5.6 Conclusion

The paradoxes arise when we apply what looks like valid rules to what looks like a perfectly meaningful sentence. Plan A solutions try to say where this reasoning goes wrong. One of the rules, or the sentence, will have to go. Plan B solutions deny this. They accept the rules as well as the sentence, but avoid the Great Collapse nonetheless. This is what we do when we pre-philosophically encounter the paradoxes, but it is not clear how it can be anything but irrational. The default conception of deductive logic is a way in which the ordinary reaction to the paradoxes can be seen as perfectly rational, and how a plan B solution to the paradoxes is possible. It has in its favor that it captures the wisdom in the natural reaction to the paradoxes, that it doesn't seem to be threatened by revenge paradoxes, and that it explicitly affirms out front

[9] The generic conception of schemas gives a nice contrast to the 'openendedness' of schemas advanced by various philosophers. According to their conception, to accept a schema is to accept any meaningful instance, expressible in our language or not. Thus if I increase my vocabulary I thereby accept another instance. This, I take it, is correct, subtleties aside. But on the generic conception one might accept a schema without accepting every instance of it. The generic conception and the openendedness conception are not incompatible, as long as the acceptance of inexpressible instances is understood generically.

what otherwise seems like a counterintuitive consequence, namely that our inference rules are not (strictly) truth preserving. They are not truth preserving or valid in the strict sense that the ideal of deductive logic hoped for, but they are truth preserving in the generic sense. Thus on this conception of validity all the rules of classical logic and the natural rules for the truth predicate are valid, generically, while at the same time some instances of them lead from truth to falsity. And while it is rational to accept the rules and to reason in accordance with them, it is also rational to reject a particular conclusion that can be drawn this way. This, it seems to me, is the answer to the philosophical problem that the semantic paradoxes pose. All this does not take away from the value in the sophisticated work that has been done in seeing which ones of the strictly valid rules avoid triviality or contradictions. But in the end it won't solve the philosophical problem with the paradoxes. What made them problematic is the ideal of deductive logic as the paradigm case of good reasoning. It is not which inference rules we took to be valid that caused the trouble, but what we took a valid inference rule to be.[10]

References

Carlson, G. N., and Pelletier, F. J., (eds.) (1995) *The Generic Book*. University of Chicago Press

Field, H. (2003), 'A revenge-immune solution to the semantic paradoxes'. *Journal of Philosophical Logic* 32: 139–177

_____ (2006a). 'Solving the paradoxes, escaping revenge'. Unpublished manuscript

_____ (2006b). 'Truth and the unprovability of consistency'. Unpublished manuscript

Friedman, H., and Sheard, M. (1987). 'An axiomatic approach to self-referential truth', *Annals of Pure and Applied Logic* 33: 1–21

Hofweber, T. (2007). 'The ideal of deductive logic'. Unpublished manuscript

Lycan, W. G. (1993). MPP, RIP. In J. Tomberlin (ed.), *Philosophical Perspectives: Language and Logic*, vol. vii. Ridgeview Publishing

_____ (2001). *Real Conditionals*. Oxford University Press

McGee, V. (1985). 'A counterexample to modus ponens', *Journal of Philosophy* 82(9): 462–71

_____ (1991). *Truth, Vagueness, and Paradox*. Hackett

Pelletier, F. J., and Asher, N. (1997). 'Generics and defaults'. In J. van Benthem, and A. ter Meulen (eds.), *Handbook of Logic and Language*. Elsevier

Priest, G. (2006). *Doubt Truth to be a Liar*. Oxford University Press

Schiffer, S. (2003). *The Things we Mean*. Oxford University Press

Simmons, K. (1993). *Universality and the Liar*. Cambridge University Press

[10] I am indebted to Keith Simmons, Bill Lycan, Graham Priest, Dean Pettit, Hartry Field, and Marc Lange for helpful discussions of this material.

6

On the Metatheory of Field's 'Solving the Paradoxes, Escaping Revenge'

Hannes Leitgeb

6.1 Introduction

In a sequence of recent publications ([3],[4],[5],[6],[2]), Hartry Field has developed a new and sophisticated account of how to approach the problem of semantic paradoxes. The core of his theory is a theory of truth for languages that contain their own truth predicate, but Field has also shown how analogous type-free theories for satisfaction ([2]), property instantiation ([5]), and definability ([2]) can be introduced. Moreover, the theory can perhaps also be used to give consistent syntactic treatments of modalities in terms of modal predicates of sentences (cf. section 9 on designatedness in [2]). Field's theory (together with related ones such as Yablo[12]) is now one of the main contenders in this area of philosophical logic and deserves critical attention from both philosophers and logicians who—like Field—are prepared to give up classical logic in view of the benefits from accepting Tarski's T-scheme unrestrictedly, as well as from those who are willing to abandon the full strength of the T-scheme while not constraining the laws or rules of classical reasoning (I count myself among the latter).

There are no fast and easy decisions to be made with regard to these alternatives: first of all, both classical logic and the T-scheme can be restricted in various different ways and often the advantages of one theory over another depend crucially on the details of these differences; e.g. Field's theory is put to the test by theories of truth that also support the unrestricted T-scheme but which, in contrast to Field, allow for sentences of the form $\varphi \wedge \neg\varphi$ to be logically acceptable (see e.g. Priest's criticism of Field's theory in his [10]). Secondly, the contenders may not even agree on what precisely their theories give up or preserve. E.g. the defenders of classical logic might complain that is not really the *original* T-scheme that holds unrestrictedly in a non-classical theory such as Field's: by weakening the logic of implication and hence of equivalence in at least some linguistic contexts, an equivalence of form T for a sentence φ that creates such a context does not have the same logical strength as the original material equivalence of form T would have had. On the other hand, non-classical truth theorists might question whether 'logic' actually remains unaffected once instances of the T-scheme are declined: if, as some deflationists have it, the truth predicate is ascribed the status of a 'quasi-logical' constant, then the denial of some of its most basic and intuitively plausible laws counts as a 'quasi-logical' theory change the status of which differs from, say, the revision of empirical hypotheses against the background of an otherwise stable logical framework.

Unfortunately, we still lack a satisfying methodology by which the merits or shortcomings of theories that are governed by different systems of logic could be assessed comparatively; this topic should receive increased attention in future discussions on theories of truth for semantically closed languages. In the meantime, we have to help ourselves by at least getting a clear understanding of what the logical and philosophical consequences are that the acceptance of any of these theories would commit us to, and what the presuppositions are on which the acceptance of any such theory rests. My following remarks on Field's theory are thus best understood as an invitation to fill in some of the blanks that have been left open by Field's own discussion of the theory. I see the main such incompleteness in the lack of an *explicitly stated metatheory of his theory of truth*, in a sense that I am going to explain below. Such a metatheory can be regarded as separate from Field's theory of truth, and its language level is indeed likely to differ from the level of the latter. But Field might just as well be able to show that such a metatheory can be introduced on the same linguistic level as his theory of truth. Perhaps he can even derive this metatheory from the resources of a mature version of his theory of truth itself. In any case, something along the lines suggested below has to be developed into a systematic metatheoretic account or otherwise the intended interpretation of Field's theory of truth, and thus his solution to the semantic paradoxes, remains unclear. Moreover, separating the (explicitly or implicitly stated) principles in 'Solving the Paradoxes, Escaping Revenge'

into *theoretical* and *metatheoretical* ones proves to be useful in order to assess Field's account properly.

I will sketch in section 6.4 what I think is demanded of such a metatheoretical interpretation of Field's theory: the special character of this metatheory is due to the fact that it must include a non-classical set theory the logic of which is identical to the logic of Field's theory of truth. Section 6.3 is devoted to an outline of what can be regarded as a *classical* metatheory of Field's theory of truth. In the context of this chapter, this metatheory is of merely pedagogical and expository value: comparing the one metatheory to the other will be helpful to understand the specific properties of the non-classical metatheory that I speculate about in section 6.4. It is only the latter that conveys what Field's actual intended interpretation of his theory of truth looks like or should look like. The classical metatheory is as close as one can get to this intended interpretation while still accepting classical logic.

In section 6.5 I will discuss the two metatheories and I will try to show that although none of them is stated explicitly in Field's paper, he nevertheless speaks at certain places *as if* he interpreted his theory of truth from the background of either of these metatheories. The fact that one part of Field's theory—the definition of logical consequence—seems to mix the two metatheoretic viewpoints obscures the actual properties of the intended models of his theory of truth. An explicit development of the intended interpretation of his theory in terms of a non-classical metatheory should help to clarify what the adoption of Field's theory of truth would ultimately commit us to—semantically, epistemically, and ontologically.

But before I turn to the two metatheories, let me highlight the main features of what I take to be Field's theory of truth. I will follow Field's development of the theory in his 'Solving the Paradoxes, Escaping Revenge', except that I add the truth predicate to the first-order language of *arithmetic* rather than to the first-order language of set theory; this has certain methodological advantages as will be pointed out later.

6.2 Field's Theory of Truth

Field's theory of truth treats the truth predicate *Tr* as a *primitive* predicate, i.e. *Tr* is not explicitly defined but rather axiomatized by this theory.

The theory can be reconstructed as follows:

1. Logical vocabulary:
   - standard logical signs of first-order languages
     (including an implication sign $\supset$ that is defined by: $\varphi \supset \psi =_{df} \neg\varphi \vee \psi$)
   - an additional distinguished implication sign $\rightarrow$

2. Descriptive vocabulary:
   - truth predicate Tr
   - the signs of first-order arithmetic, i.e. predicates for addition ($+$), multiplication ($\cdot$), and order ($<$) of natural numbers, as well as individual constants 0 and 1
   - for every formula φ of the language, an individual term $\ulcorner\varphi\urcorner$ is available that can be used as the standard name of this formula
   - (optional: further descriptive signs; e.g. special vocabulary for empirical theories)
3. Logic:
   - logical axioms and rules for Strong Kleene logic (regarding the first-order logical vocabulary)
   - additional logical axioms and rules for $\rightarrow$, according to which $\rightarrow$ turns out to be a weakening of the material implication sign of classical logic (for every axiom or rule for $\rightarrow$ there is a corresponding axiom or rule for material implication but not the other way round)
4. Eigenaxioms:
   - all instances of the T-scheme (formulated with $\leftrightarrow$, where $\leftrightarrow$ is defined in terms of $\rightarrow$ and $\wedge$ in the standard manner)
   - a sequence of definitions of sentential 'determinately-true' operators D, D^2, D^3, ... D^α, ... (for a sequence of ordinals α below some fixed ordinal that can be characterized metatheoretically); in particular, $D\varphi$ is defined as $\varphi \wedge \neg(\varphi \rightarrow \neg\varphi)$ (or alternatively as $\varphi \wedge (\top \rightarrow \varphi)$). I skip the technical intricacies (see sections 14–19 in [2])
   - an axiomatic system of arithmetic: e.g. Peano arithmetic (but at least Robinson arithmetic should be contained in it).

Discussion

As far as the language of Field's theory of truth in [2] is concerned, the only slightly surprising fact about it is his choice of including the first-order language of set theory as a sublanguage. More usually, truth theorists add the truth predicate to the vocabulary of first-order arithmetic with the intention of employing arithmetic as the background theory of syntax (modulo coding). While having the membership predicate and a version of set theory at his disposal certainly increases the expressiveness of the language of Field's theory of truth, combining set theoretic issues with truth theoretic ones also makes it more difficult to uphold Ramsey's distinction of semantic paradoxes vs. the classic set theoretic ones. In [5] Field seems to accept the now standard response to the latter in terms of axiomatic set theories such as ZFC, so the intention behind his choice of language does not seem to be to attack both kinds of paradoxes at the same time. I wonder whether adding Tr to the arithmetical vocabulary might thus have been the better choice, at least from an expository point of view. E.g. set theoretic facts such as that quantifiers in set theoretic models can only range over sets and thus not over the universe of all sets—as discussed by Field

in section 4.8—would not even have come up if he had employed arithmetic as his formal background theory. Therefore, I use an axiomatic system of arithmetic rather than set theory as the base theory that is extended by Field's theory of truth. I assume Field regarded the inclusion of set theory as necessary in order to illuminate the claimed revenge-immunity of his theory of truth. However, I think the proper place to address this topic would rather have been a metalinguistic account such as the one that I will outline in section 6.4—I will return to this topic further below.

What can be said about the logic of Field's theory? Strong Kleene logic is a well-known and useful system of partial logic; its main attraction or drawback—depending on one's aims and background assumptions—is the absence of the excluded middle from its set of logical axiom schemes. Extending Strong Kleene logic by three-valued conditional signs such as the Lukasiewicz implication leads to logics that can be regarded as the three-valued restrictions of fuzzy logic (see Hajek *et al.* [7] for a study of combinations of the latter with a theory of type-free truth); however, this is *not* the road taken by Field. The set of logical laws for his additional implication sign $\rightarrow$ are not valid with respect to any three-valued truth table (nor, for that matter, with respect to any finite-valued truth table as I suppose). What Field's implication *does* share with the logical signs of various three-valued logics is the property of extending classical first-order logic conservatively. In particular, if the excluded middle is derivable for a sentence φ, then $\varphi \rightarrow \psi$ and $\varphi \supset \psi$ are interderivable. If additionally the excluded middle for ψ can be derived, then the equivalence $(\varphi \supset \psi) \leftrightarrow (\varphi \rightarrow \psi)$ is derivable as well. The main weakness of Field's conditional is its lack of satisfying the contraction axiom scheme: accordingly, $(\varphi \wedge (\varphi \rightarrow \psi)) \rightarrow \psi$ is not valid (though of course the *rule* of modus ponens is valid). For the same reason, there is no deduction theorem for $\rightarrow$, nor is it the case that φ can be derived from $\top \rightarrow \varphi$. Semantically, these non-standard features of Field's implication operator show up in its neighbourhood semantics, which is developed in [4]: a neighbourhood semantics is itself usually chosen in order to prove logics sound and complete that are supposed to be weaker than logics that are sound and complete with respect to a possible worlds semantics; furthermore, in Field's semantics, worlds correspond to Strong Kleene truth-value assignments rather than to classical ones and Field also has to allow for 'abnormal' worlds which do not happen to be members of all the neighbourhood sets that surround them.

Are these features of Field's conditional sign sufficient reason for not accepting his theory of truth since the latter depends crucially on the new implication symbol? *No*—one does not need to rely on Quine's metaphor of a holistic web of belief in order to point to the fact that if a theory as a whole is to be judged, the properties of a single part of it are ultimately irrelevant if taken in isolation from the other parts of the theory. Indeed, every 'solution' to the problem of semantic

paradoxes must face the task of shifting the contradictory consequences of the latter to a 'place' where they show up as merely unintended though consistent properties of our theoretical treatment. Abandoning the excluded middle and the contraction axiom scheme might just be the right location at which our prima-facie theories of truth have to be revised—who knows? In the context of Field's theory of truth, the absence of the principle of excluded middle shows up more or less positively, i.e. as a manner of blocking the derivation of contradictions from the T-instance for the Liar sentence. The lack of contraction has mainly the effect that Field is forced to state some of his axioms and theorems by means of the defined relation of logical consequence rather than in terms of $\rightarrow$ (see pp. 14, 20, 43, 50, 51, 52 of [2]); this is inconvenient but by itself certainly not a reason for rejecting the theory.

On the other hand, Field's claim that this change of logic would only affect the semantical parts of our scientific theories but that it would leave their mathematical and empirical parts perfectly intact has to be taken with caution: it is true that our standard mathematical-semantical bridge principles such as

$$\forall x(Prov(x) \rightarrow Tr(x))$$

(where '*Prov*' expresses mathematical provability) would only be affected if we were forced to abandon the excluded middle for *mathematical* sentences, which is not the case according to Field's theory. However, it is not clear why there *could* not be mathematical-semantical bridge principles that *would* be compromised by dropping the excluded middle as a general logical axiom for semantic sentences. E.g. consider the claim that a proper theory of truth has to be conservative over mathematics and assume that Field's theory of truth turned out to be *not* conservative in this sense: the fact that a certain new mathematical theorem—rather than, say, its negation—would be derivable from the combination of Field's theory of truth with a mathematical theory might actually hinge on Field's revision of certain logical laws; the logic of truth thus could not be changed without a corresponding revision of our mathematical theories.

As far as the logical differences of semantical and empirical theories are concerned, Field's claim seems to entail that any physicalistic reduction of semantic concepts to empirical ones is implausible if not entirely impossible. What would the Field of 1972's 'Tarski's Theory of Truth' say about these prospects? The clear-cut separability of logical restrictions for semantic concepts and logically unrestricted reasoning for non-semantic concepts is thus neither obvious nor unproblematic.

Let us turn now to the actual *truth*-theoretic part of Field's theory. The acceptance of the unrestricted T-scheme and the related general intersubstitutivity result for $Tr(\ulcorner\varphi\urcorner)$ and φ constitute highly attractive features of Field's theory of truth and

represent its most important progress over Kripke's [8] theory: while according to Kripke the sentences $Tr(\ulcorner\varphi\urcorner)$ and φ indeed have the same semantic value, their being equivalent cannot be expressed on the object-linguistic level due to the lack of a corresponding equivalence sign; not so in Field's theory.

Since Field is well aware that for some purposes an additional unrestricted *satisfaction* rather than a *truth* scheme is needed, he is happy to endorse the former as well. If the satisfaction scheme still turns out to be too weak in order to derive all the general truths about truth from it that we are interested in (which I suppose to be the case), Field can still add to the deductive power of his theory by introducing further axioms for *Tr*. I will exclude any further discussion of this more general type of problem that disquotational theories of truth have to cope with. Moreover, I am going to postpone the discussion of the sequence of defined 'determinately-true' operators, since their role will get much clearer once the classical metatheory of Field's theory of truth has been introduced in the next section.

As said above, I include an axiomatic system of arithmetic as a subtheory of Field's theory of truth; this arithmetical component can be used as a theory of syntax via assigning natural numbers codes to sentences effectively.

Now that I have stated and explained Field's theory of truth, I may thus turn to the question of how and whether this theory can be defended.

Any justification of a theory *T* must be stated in terms of a theory *MT* that speaks *about T*, therefore I refer to such a theory *MT* as a *metatheory* of *T*. A weak but nevertheless very important form of justification is given in terms of a metalinguistic consistency proof of *T*: as far as Field's theory of truth is concerned, this has been achieved by Field in his [3] on the basis of an ingenious model construction that combined Kripke's transfinite Strong Kleene truth approximation with Herzberger's, Gupta's, and Belnap's revision sequences of truth extensions. However, a consistency proof is by itself not *sufficient* for the justification of a theory (except perhaps in the case of mathematical theories). In order to prove a theory consistent we are allowed to use any interpretation of the primitive signs of the theory we like, as long as all the members of the theory turn out to be true under that interpretation. For any stronger form of justification the interpretation of the theory's vocabulary must be narrowed down to one *intended* interpretation (the standard model of the theory) or at least to a *set* of such intended interpretations. Field made it very clear that the model in [3] is not among the latter; no 'meaning' whatsoever is assigned to it. Accordingly, neither Kripke's notion of groundedness, nor the revision understanding of T-biconditionals as partial circular definitions of truth play any role in Field's own interpretation of his theory. Furthermore, if the model in [3] were actually taken to *define* Field's theory of truth rather than merely to prove it consistent, then a huge lot of set theoretic machinery would automatically be imported into Field's concept of truth. As shown by Philip Welch ([11]), this would

leave Field with a recursion-theoretically highly complex (overly complex?) theory of truth.

Fortunately, Field can avoid these problems by separating the metatheoretic justification of his theory of truth from the specific details of this particular set theoretic model. As I understand it, a good part of 'Solving the Paradoxes, Escaping Revenge' is about filling the justificational gap thus left by [3] while at the same time generalizing the theory of truth that was presented in [3] to a *class* of theories which are called 'G-solutions'. A G-solution is a theory of truth that (i) includes the full T-schema expressed in terms of an implication sign $\rightarrow$ and (ii) the logic of which neither includes the law of excluded middle nor the intersubstitutivity of $\varphi \rightarrow \psi$ and $\neg\varphi \vee \psi$. In short, a G-solution is a theory *like* the theory that I dealt with in this section, except that the latter also includes a system of arithmetic as a proper fragment. In the next two sections I will consider two possible metatheories of this latter theory, but practically all that is said by these metatheories actually applies to *any* G-solution. The purpose of the metatheories is to describe and thereby constrain the intended models of Field's theory of truth. In section 6.3 these intended models are described in a *classical* metatheory; if anything, they are intended from the point of view of classical logicians but certainly not according to Field's standards. What I consider to be Field's own implicit interpretation of his theory is made explicit in section 6.4, however, our lacking a clear account of the kind of non-classical set theory that would be demanded by such a metatheory leaves us—or rather Field—with an open problem. In order to specify the axioms and definitions of both the classical and the non-classical metatheory, I use hints given by Field in his [3].

One final remark about the status of these metatheories: it would be a mistake to downplay their importance on the basis of worries concerning *circularity* or *triviality* or *regress*: first of all, there is nothing circular or trivial about interpreting a non-classical theory of truth by means of a classical metatheory. Secondly, there is nothing *viciously* circular about interpreting a non-classical theory of truth in terms of a non-classical metatheory (nor is there a problem about deriving the existence of the classical model of arithmetic within classical set theory). The aim of the non-classical metatheory is to make the intended non-classical understanding of Field's non-classical theory of truth more explicit; the method of achieving this is by actually stating what the symbols in Field's theory are supposed to refer to. Some of the tacit assumptions and presuppositions of Field's theory will show up as definite axioms and theorems in the metatheory. The revenge-immunity of Field's theory of truth should be derivable from these axioms and theorems. If for some reason such a non-classical metatheory could *not* be developed, this would speak *against* Field's theory of truth; the possibility of this being the case shows that there is nothing trivial about the introduction of such metatheories. Hence it must be in the interest of both supporters and critics of

Field's theory to review the prospects of metatheoretic accounts as the ones that I will turn to now.

6.3 A Classical Metatheory of Field's Theory of Truth

The following theory is supposed to describe what a member of the set of intended interpretations of Field's theory of truth is like from a classical point of view. This intended interpretation is not singled out uniquely by this description, but the latter at least enumerates various necessary constraints that must be satisfied by *any* intended classical interpretation of Field's theory.

The theory is structured as follows:

1. Logical vocabulary:
   - standard logical signs of first-order languages
2. Descriptive vocabulary:
   - membership predicate $\in$
   - for every formula φ of the language of Field's theory of truth, an individual term $\ulcorner\varphi\urcorner$ is available that can be used as the standard name of this formula
   - a finite sequence of individual constants (which are intended to denote the different components of an intended interpretation of Field's theory of truth): $\mathsf{DM}, \leq, \sim, \mathsf{1}, \mathsf{0}, \mathsf{D}, \mathsf{I}, \mathsf{Val}, \mathsf{P}, \ldots$
   - (optional: further descriptive signs; e.g. special vocabulary for empirical theories)
3. Logic:
   - standard first-order logic
4. Eigenaxioms:
   - ZFC set theory (if there is empirical vocabulary, a version of ZFC with urelements has to be used; if ZFC is regarded as too strong, ZF may be used alternatively)
   - definitions and axioms that are used to describe the components of an intended model of Field's theory:

Instead of formalizing these definitions and axioms properly, I rather state them in a rather informal and sketchy way: e.g. when I postulate 'the complement function $\sim$ is a dual automorphism of period two' this is short for an axiom for the primitive sign '$\sim$' that could be made precise by first defining an auxiliary predicate P ('is a dual automorphism of period two') in terms of $\in$ and then stating as axiom:

$$P(\sim)$$

Here are the informally stated definitions and axioms of the theory:

Algebra: $\mathsf{DM} = \langle \mathsf{A}, \leq, \sim, \mathsf{1}, \mathsf{0} \rangle$ is a De Morgan algebra, i.e. a bounded distributive lattice with a complement function $\sim$ that is a dual automorphism of period two, where $\mathsf{1}$ is the top element of DM and $\mathsf{0}$ is its a bottom element. By the definition of a De Morgan algebra, the laws of double complement, contraposition, and De Morgan hold in DM. Furthermore, DM is supposed to be infinite and its underlying partial order $\leq$ is assumed to be not total. Finally, for all finite subsets of A the existence of an infimum and a supremum is postulated, for certain infinite subsets of A the existence of an infimum and a supremum is postulated, and the supremum of certain finite or infinite subsets of A is assumed to be $\mathsf{1}$ if and only if $\mathsf{1}$ is a member of this set; I say more about this in the discussion below. DM is taken to the be the co-domain of an intended valuation function on the language of Field's theory of truth.

Domain: D is a set that is used as the intended domain or universe of discourse of Field's theory of truth. D can be postulated to be identical to the set of natural numbers; it also contains all sentences of the language of Field's theory modulo (arithmetical) coding.

Interpretation: The interpretation mapping I assigns to each primitive term in the language of Field's theory of truth its *intended interpretation*. The interpretation of a standard name of a sentence φ is simply φ. The interpretation of *Tr* is a mapping from D to A where every non-sentence (or rather: everything that is not the code of a sentence) is mapped to $\mathsf{0}$. The interpretation of an n-ary arithmetical predicate is a mapping from D^n to A. All optional additional predicates are also assigned functions of the right set theoretic type; every optional additional individual constant gets assigned a member of D as its interpretation (if there is additional vocabulary, D might have to be extended accordingly). The arithmetical individual constants receive their intended interpretation.

Valuation: The valuation mapping Val is a function that assigns to each formula in the language of Field's theory of truth an element of DM (relative to variable assignments s that map variables into D; in the case of sentences I omit the reference to variable assignments). Val_S integrates the components from above into a *De Morgan-valued model* (see Leitgeb [9] for the application of such models in the context of theories of truth). On the first-order logical vocabulary, Val_S is defined just as expected, i.e. every first-order logical sign 'translates' into its intended algebraic counterpart. Concerning the additional distinguished implication sign $\rightarrow$, $\mathsf{Val}_\mathsf{S}(\varphi \rightarrow \psi)$ is assumed to be equal to $\mathsf{1}$ if and only if $\mathsf{Val}_\mathsf{S}(\varphi) \leq \mathsf{Val}(\psi)$; some further constraints on the values of conditional formulas are postulated axiomatically (see Field's list on

p.18 of [2] which is translatable into a list of additional axioms). As far as truth is concerned, $\mathsf{Val}(Tr(\ulcorner\varphi\urcorner)) = \mathsf{Val}(\varphi)$ is assumed for every sentence φ in the language of Field's theory of truth; since the value of atomic formulas is additionally supposed to be given by corresponding applications of the interpretation mapping, $\mathsf{I}(Tr)(\mathsf{I}(\ulcorner\varphi\urcorner)) = \mathsf{I}(Tr)(\varphi) = \mathsf{Val}(\varphi)$ is derivable. Finally, it is postulated that atomic arithmetical formulas receive the values $\mathsf{1}$ or $\mathsf{0}$ just as expected: e.g. $\mathsf{Val_S}(+(x, y, z)) = \mathsf{1}$ iff $\mathsf{s}(x) + \mathsf{s}(y) = \mathsf{s}(z)$ and $\mathsf{Val_S}(+(x, y, z)) = \mathsf{0}$ otherwise.

Prime filter: P is a prime filter of DM, i.e. P is a subset of A such that (i) $\mathsf{1} \in \mathsf{P}$, (ii) $\mathsf{0} \notin \mathsf{P}$, (iii) if $\mathsf{x} \in \mathsf{P}$ and $\mathsf{x} \leq \mathsf{y}$ then $\mathsf{y} \in \mathsf{P}$, (iv) $\mathsf{inf\{x,y\}} \in \mathsf{P}$ iff $\mathsf{x} \in \mathsf{P}$ and $\mathsf{y} \in \mathsf{P}$, (v) $\mathsf{sup\{x,y\}} \in \mathsf{P}$ iff $\mathsf{x} \in \mathsf{P}$ or $\mathsf{y} \in \mathsf{P}$. Moreover P is assumed to have the property that there is no $\mathsf{x} \in \mathsf{A}$ such that both x and $\sim \mathsf{x}$ are members of P, and P is also supposed to satisfy closure conditions like (iv) and (v) with respect to certain infinite infima and suprema. Further axioms could, and should, be introduced which constrain P to be a *particular* such prime filter rather than any such prime filter whatsoever. (Field [2] does not consider such prime filters himself but he acknowledges the possibility of doing so.)

Logical consequence: On the basis of the ZFC membership sign $\in$ a logical consequence relation $\vdash$ is defined that holds between *sets* of formulas in the language of Field's theory of truth and *formulas* of this language. The definition of logical consequence will be in terms of truth preservation in all G-models and thus will involve a qualification over all G-models. I won't go into any details, but roughly a *G-model* is defined to be any set theoretic model for this language (i) the domain of which is the set of natural numbers and (ii) for which conditions analogous to the ones that I have stated for the intended model are satisfied: where I used primitive terms such as DM, D, and Val in order to denote the components of the particular intended model that I focused on, 'G-model' is defined by replacing each of these primitive terms by a variable and by turning each of our original axioms into a definitional clause.

A formula in the language of Field's theory of truth is called logically true if and only if it is a consequence of the empty set (for convenience, arithmetical truths will thus be counted as logical truths—just as Field includes set theoretic truths in the set of logical truths in [2]).

Determinate truth: A sentence in the language of Field's theory of truth is defined to be determinately true if and only if Val maps this sentence to $\mathsf{1}$, i.e.:

$$\forall x(\mathsf{DetTr}(x) \text{ iff } \mathsf{Val}(x) = 1)$$

Truth: Finally, a sentence in the language of Fields's theory of truth is defined to be true if and only if Val maps this sentence into P, i.e.:

$$\forall x(\mathsf{Tr}(x) \text{ iff } \mathsf{Val}(x) \in \mathsf{P})$$

Discussion

The members of $\mathbf{DM}$ (or rather $\mathbf{A}$) are used as the semantic values of formulas in the language of Field's theory of truth. What is the proper interpretation of these values? It is tempting to consider them as *credences*, i.e. as numerical degrees of belief in the closed real interval [0,1], however this interpretation is excluded by the axiomatic postulation of a *non-total* partial order $\leq$. Furthermore, Field explicitly denies any *epistemic* account of truth value assignments in [4] (this does not mean that there could not be a non-classical theory of probability that would stand to classical probability theory as Field's theory of truth stands to classical theories of truth; cf. section 3 in [4]). $\mathbf{1}$ is rather to be regarded as the *determinately-true*-truth-value, $\mathbf{0}$ as the *determinately-false*-truth-value, and every other value as something like a non-numerical degree of truth-determinacy that is located somewhere in between $\mathbf{1}$ and $\mathbf{0}$.

Since Field accepts the Strong Kleene rules of disjunction elimination and existential elimination in his theory of truth, the metatheoretical definition of the logical consequence relation $\vdash$ must be such that these two rules turn out to be valid. Every intended classical model of the theory should be among the models in terms of which logical consequence is defined, therefore additional axioms are needed which state that the supremum of certain finite or infinite subsets $\mathbf{X}$ of $\mathbf{A}$ is identical to $\mathbf{1}$ if and only if $\mathbf{1}$ is a member of $\mathbf{X}$ (if this is the case for $\mathbf{X}$, I will say that $\mathbf{X}$ has the supremum property). Why not simply postulate the supremum property for *arbitrary* subsets of $\mathbf{A}$? As Field notes on p.17 of [2], this would exclude algebras $\mathbf{DM}'$ the fields $\mathbf{A}'$ of which have the property that $\mathbf{A}' \setminus \{\mathbf{1}\}$ has no largest element: otherwise $\mathbf{sup}(\mathbf{A}' \setminus \{\mathbf{1}\}) = \mathbf{1}$ though $\mathbf{1} \notin \mathbf{A}' \setminus \{\mathbf{1}\}$. But it is precisely algebras of this latter form that play an important role in Field's intended interpretation of his theory of truth. Therefore, at least $\mathbf{A}' \setminus \{\mathbf{1}\}$ should not be assumed to have the supremum property. So which subsets of $\mathbf{A}$ *shall* be postulated to have it?

It would actually suffice to assume the supremum property for all those sets $\mathbf{X}$ of values in $\mathbf{A}$ that can be expressed in terms of formulas in the language of Field's theory of truth, i.e. where for some such formula φ, for some free variable x in φ, and for some variable assignment $\mathbf{s}$ such that $\mathbf{S}$ is the set of all variable assignments that differ from $\mathbf{s}$ at most at x, it holds that $\mathbf{X} = \{\mathbf{x} \in \mathbf{A} | \exists \mathbf{s}' \in \mathbf{S} \text{such that} \mathbf{x} = \mathbf{Val}_{\mathbf{s}'}(\varphi)\}$. $\mathbf{A}' \setminus \{\mathbf{1}\}$ would thus not be postulated to have the supremum property as long as it did not figure as a set of values in $\mathbf{DM}$ that is expressed by a formula in the language of Field's theory of truth. However, Field proceeds in a different way in [2]: in order to formulate a strictly algebraic constraint on algebras, i.e. a constraint that is not formulated in terms of $\mathbf{Val}$, he suggests to demand the supremum property for all sets $\mathbf{X} \subseteq \mathbf{A}$ the cardinality $|\mathbf{X}|$ of which is less than or equal to the cardinality of the

domain D. This postulate has the following consequence: since A is assumed to be infinite, if the cardinality of A were less than or equal to to the cardinality of D then the same would hold for $A' \setminus \{1\}$; thus the supremum property would apply to it and entail a contradiction as explained above. Therefore, |A| must be greater than |D|. This seems to invite non-trivial formal and philosophical questions concerning what would be the case if |D| were not a set but rather identical to the *class of all sets*: how could the cardinality of A be larger than the cardinality of the set-theoretic universe (given it were meaningful to speak of the cardinality of the universe of sets at all)? In my view, a discussion like this would mean to exaggerate this postulate of Field's: as Field has shown himself, it is possible to apply the methods in [3] to construct a model for a G-solution that is formulated in the language of first-order arithmetic extended by the truth predicate. The intended domain of this model is countably infinite; the algebra of the model can be restricted to the countably infinite set of values that the sentences of the language have in this model. Although the domain and the algebra of the resulting model thus have the same cardinality, no problem does arise: the existence and the properties of the constructed model imply that there is no set $X \subseteq A$ that is expressible by a formula and which at the same time fails to have the supremum property. Thus, Field's postulate from above is unnecessarily strong and hence should not be taken to be of crucial importance. Accordingly, also the *existence* of infima and suprema and the closure properties of P with respect to infima and suprema ought to be postulated for sets of values that can be expressed by formulas rather than for all sets of a cardinality less than or equal to the cardinality of the domain.

The constraints on the evaluation of atomic formulas that include *Tr* or the arithmetical signs are such that the T-biconditionals and the arithmetical axioms in Field's theory of truth are satisfied. Hence, it is possible to actually derive in our classical metatheory that Val assigns to every theorem in Field's theory of truth the semantic value 1. Therefore, ever theorem of Field's theory can be shown DetTr and thus Tr in our metatheory. In this sense the intended model indeed proves to be a model of Field's theory.

Furthermore, the metatheory may be used to address *revenge* type problems, i.e. to answer questions such as: what kinds of facts about the components of the intended model can be expressed on the basis of terms or formulas in the language of Field's theory of truth as interpreted by this model? Let us concentrate on three of these components: the partial order $\leq$ of the intended algebra, the set $\{\varphi \mid \mathsf{DetTr}(\varphi)\}$ of formulas that are determinately true, and the set $\{\varphi \mid \mathsf{Tr}(\varphi)\}$ of formulas that are true as being given by the intended model.

E.g.: it is a set theoretic fact that there exists a function F that assigns to every pair of formulas in the language of Field's theory of truth either 1 or 0 according to the

following definition:

$$\bullet\ \mathsf{F}(\varphi, \psi) = 1 \text{ iff } \mathsf{Val}(\varphi) \leq \mathsf{Val}(\psi)$$
$$\bullet\ \mathsf{F}(\varphi, \psi) = 0 \text{ iff } \mathsf{Val}(\varphi) \not\leq \mathsf{Val}(\psi)$$

Note that since the metatheory is classical, the 'iff' is simply the standard material equivalence sign and the following instance of the excluded middle can be derived:

$$\bullet\ \mathsf{Val}(\varphi) \leq \mathsf{Val}(\psi) \vee \mathsf{Val}(\varphi) \not\leq \mathsf{Val}(\psi)$$

Is it possible to express this 'semantic' function F in terms of a binary predicate that is definable in the language of Field's theory of truth, i.e. is there a predicate *Imp* for which

Imp1: $\mathsf{Val}(\mathit{Imp}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner)) = 1 \text{ iff } \mathsf{Val}(\varphi) \leq \mathsf{Val}(\psi)$
Imp2: $\mathsf{Val}(\mathit{Imp}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner)) = 0 \text{ iff } \mathsf{Val}(\varphi) \not\leq \mathsf{Val}(\psi)$

is derivable in the metatheory? It turns out that this is not the case, for otherwise: (by some method of diagonalization) there would a sentence ρ in the language of Field's theory of truth for which

$$\bullet\ \rho \leftrightarrow \mathit{Imp}(\ulcorner\rho\urcorner, \ulcorner\bot\urcorner)$$

could be derived in Field's theory; accordingly, we would be able to derive

$$\bullet\ \mathsf{Val}(\rho \leftrightarrow \mathit{Imp}(\ulcorner\rho\urcorner, \ulcorner\bot\urcorner)) = 1$$

and thus

$$\bullet\ \mathsf{Val}(\rho) = \mathsf{Val}(\mathit{Imp}(\ulcorner\rho\urcorner, \ulcorner\bot\urcorner))$$

in the metatheory. Since $\mathsf{Val}(\bot) = 0$ is derivable metatheoretically as well, both the assumption $\mathsf{Val}(\rho) \leq \mathsf{Val}(\bot)$ and $\mathsf{Val}(\rho) \not\leq \mathsf{Val}(\bot)$ lead to a contradiction (this is of course simply a version of the *Curry paradox*). Accordingly, the implication sign $\rightarrow$ in Field's theory expresses the partial order of the intended model only in the weaker sense that a version of [Imp1] is derivable for it; however, the corresponding version of [Imp2] cannot be derived, which saves the metatheory from inconsistency but which also reflects the impossibility of representing the function F in the language of Field's theory of truth completely.

Analogous considerations can be directed towards DetTr: ZFC allows us to prove that there is a function G such that

$$\bullet\ \mathsf{G}(\varphi) = 1 \text{ iff } \mathsf{DetTr}(\varphi)$$
$$\bullet\ \mathsf{G}(\varphi) = 0 \text{ iff } \neg\mathsf{DetTr}(\varphi)$$

and where of course

$$\bullet\ \mathsf{DetTr}(\varphi) \vee \neg\mathsf{DetTr}(\varphi)$$

is the case.

If a unary predicate *DetTr* were definable in the language of Field's theory of truth such that

DetTr1: $\mathsf{Val}(\mathsf{DetTr}(\ulcorner\varphi\urcorner)) = 1$ iff $\mathsf{DetTr}(\varphi)$
DetTr2: $\mathsf{Val}(\mathsf{DetTr}(\ulcorner\varphi\urcorner)) = 0$ iff $\neg\mathsf{DetTr}(\varphi)$

were derivable in the metatheory, then *DetTr* would express the 'semantic' function G. However, this turns out to be impossible again, for otherwise: (by diagonalization) there would a sentence σ in the language of Field's theory of truth for which

$$\bullet\ \sigma \leftrightarrow \neg DetTr(\ulcorner\sigma\urcorner)$$

could be derived in Field's theory; accordingly,

$$\bullet\mathsf{Val}(\tau \leftrightarrow \neg\mathsf{DetTr}(\ulcorner\sigma\urcorner)) = 1$$

and thus

$$\bullet\ \mathsf{Val}(\sigma) = \mathsf{Val}(\neg\mathsf{DetTr}(\ulcorner\sigma\urcorner))$$

would be derivable in the metatheory. But then both the assumption $\mathsf{DetTr}(\sigma)$ and $\neg\mathsf{DetTr}(\sigma)$ lead to a contradiction.

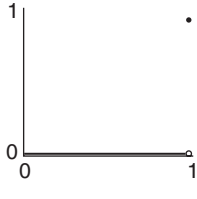
As Field shows in several of his papers, the intended notion of determinate truth can at least be approximated in terms of sentential operators

$$D, D^2, D^3, \ldots, D^\alpha, \ldots$$

that can be defined in the language of Field's theory of truth for a sequence of ordinals α below some fixed ordinal (recall section 2). Each of these operators expresses *some* of the properties of DetTr; e.g.

$$\text{if } \mathsf{DetTr}(\varphi) \text{ then } \mathsf{Val}(D^\alpha\varphi) = 1$$

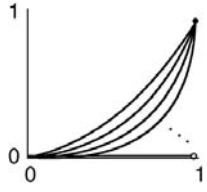
can be derived metatheoretically. The approximation of DetTr in terms of iterations of D can be visualized as follows: let us (unjustifiedly) use the compact real interval [0,1] in order to represent our set of semantic values; points on the x-axis are supposed to stand for the values $\mathsf{Val}(\varphi)$ of sentences φ in the language of Field's theory; I use the y-axis to depict the intended corresponding values of $\mathsf{G}(\varphi)$. This yields:



Thus, the value of $\mathsf{G}(\varphi)$ is $\mathsf{0}$ except for the single case where φ has value $\mathsf{1}$ in which case $\mathsf{G}(\varphi)$ has value $\mathsf{1}$ accordingly (note that any iteration of the mapping in our

diagram would be identical to the original mapping, which is why Field speaks of the *idempotency* of such 'absolute' determinacy operators).

Now I add the value functions as being given by D, D^2, D^3, . . . , D^α, . . .:



We see that these functions approximate the function G without ever actually reaching it. (Note that—though useful—these graphical representations are actually inaccurate: the partial order of the intended algebra is not total; thus Field's principle (d) for D-operators cannot be satisfied by the curves in our diagrams; see p. 30 of [2]).

Finally, *truth*: there exists a function H such that

- $\mathsf{H}(\varphi) = 1$ iff $\mathsf{Tr}(\varphi)$
- $\mathsf{H}(\varphi) = 0$ iff $\neg\mathsf{Tr}(\varphi)$

is the case with

- $\mathsf{Tr}(\varphi) \vee \neg Tr(\varphi)$

being derivable.

If a unary predicate P were definable in the language of Field's theory of truth such that

Tr1: $\mathsf{Val}(P(\ulcorner\varphi\urcorner)) = 1$ iff $\mathsf{Tr}(\varphi)$
Tr2: $\mathsf{Val}(P(\ulcorner\varphi\urcorner)) = 0$ iff $\neg\mathsf{Tr}(\varphi)$

were derivable in the metatheory, then P would express this 'semantic' function H. But this cannot be so, for otherwise: (again by diagonalization) there would a sentence τ in the language of Field's theory of truth for which

- $\tau \leftrightarrow \neg P(\ulcorner\tau\urcorner)$

could be derived in Field's theory; hence,

- $\mathsf{Val}(\tau \leftrightarrow \neg P(\ulcorner\tau\urcorner)) = 1$

and thus

- $\mathsf{Val}(\tau) = \mathsf{Val}(\neg P(\ulcorner\tau\urcorner))$

would be derivable in the metatheory. But then both the assumption $\mathsf{Tr}(\tau)$ and $\neg\mathsf{Tr}(\tau)$ lead to a contradiction. This still leaves us with the possibility of representing

some aspects of metalinguistic truth within the object language: in particular, the truth predicate *Tr* of Field's theory of truth at least satisfies the left-to-right directions of [Tr1] and [Tr2].

We found that $\leq$, DetTr, and Tr cannot be expressed completely in the language of Field's theory of truth and that our classical metatheory can be used to prove this. In the terminology of Field: we are confronted with a *revenge problem*. More precisely, Field's theory of truth is affected by a revenge problem *with respect to the classical metatheory that was developed in this section.*

One could now start to discuss how severe these revenge problems are and thereby judge the merits and shortcomings of Field's theory. Fortunately, I do not have to enter this discussion, because Field is actually *not* defending a classical interpretation of his theory of truth at all. Instead, the informal discussion of his theory of truth in [2] seems to point to a non-classical understanding of the theory in terms of the non-classical metatheory that I will turn to now.

6.4 A Non-classical Metatheory of Field's Theory of Truth

The following theory is supposed to describe Field's intended non-classical interpretation of his theory of truth. While this metatheory has not been developed by Field explicitly, at various places he speaks as if it were presupposed implicitly--we will discuss this in more detail below. I regard the precise statement of the axioms of this metatheory as an *open problem*. But at least it can be pointed out what the consequences of the theory ought to be: none of the *classical sets* such as $\leq$ or $\{\varphi|\mathsf{DetTr}(\varphi)\}$ or $\{\varphi|\mathsf{Tr}(\varphi)\}$ that figured so prominently in the metatheory of the last section is going to *exist* in the non-classical metatheory. I think this is ultimately Field's solution to the revenge problem: what showed up as *non-expressibility* in the last section is now turned into plain *non-existence*; there is no fact of the matter left on the basis of which a 'counterattack' could be launched.

The theory should consist of the following parts:

1. Logical vocabulary:
   - standard logical signs of first-order languages
   - the additional distinguished implication sign $\rightarrow$
2. Descriptive vocabulary:
   - membership predicate $\in$
   - for every formula φ of the language of Field's theory of truth, an individual term $\ulcorner\varphi\urcorner$ is available that can be used as the standard name of this formula

- a finite sequence of individual constants (which are again intended to denote the different components of an intended interpretation of Field's theory of truth): DM, $\leq$, $\sim$, $\mathsf{1}$, $\mathsf{0}$, D, I, Val, P . . .
- (optional: further descriptive signs; e.g. special vocabulary for empirical theories)

3. Logic:
   - logical axioms and rules for Strong Kleene logic (regarding the first-order logical vocabulary)
   - the logical axioms and rules that are used for $\rightarrow$ in Field's theory of truth
4. Eigenaxioms:
   - definitions of a sequence of sentential 'determinately-true' operators D, D^2, D^3, . . . , D^α, . . . as in Field's theory of truth (though perhaps extended to a longer sequence up to a higher ordinal)
   - a non-classical set theory that (i) has ZFC as a proper classical fragment, (ii) includes additional versions of separation, choice, and transfinite recursion which are restricted to formulas for which the excluded middle holds; e.g. separation is restricted in the way that

$$\forall x \in z(A[x] \vee \neg A[x]) \vdash \exists y \forall x(x \in y \leftrightarrow x \in z \wedge A[x])$$

where '$\vdash$' is defined set theoretically in the theory (see below). Finally, (iii) there is an extensionality axiom and there are existence axioms for sets the membership conditions of which are 'indeterminate$^\alpha$' for any α referred to above, i.e.: first of all,

$$\forall x \forall y(D^\alpha(y \in x) \vee D^\alpha(y \notin x))$$

and

$$\forall x \forall y D^\alpha(y \in x \vee y \notin x)$$

are *not* derivable. If we include the case $\alpha = 0$ in the latter (with $D^0\varphi = \varphi$), this means that also

$$\forall x \forall y(y \in x \vee y \notin x)$$

will not be derivable. Indeed it should be the case that, say,

$$\exists x \exists y(\neg D^\alpha(y \in x) \wedge \neg D^\alpha(y \notin x))$$

is derivable from the axioms, or equivalently

$$\exists x \exists y \neg(D^\alpha(y \in x) \vee D^\alpha(y \notin x))$$

(note that fuzzy set theory in the standard sense could not be employed for this purpose, due to the crucial differences between fuzzy logic and Field's non-classical logic that

were discussed in section 2). One of the general and most interesting properties of this non-classical set theory will be that a 'classical set' can have 'non-classical sets'—i.e. sets with indeterminate$^{\alpha}$ membership conditions—as subsets

- definitions and axioms that are used to describe the non-classical set theoretic components of a particular intended model of Field's theory:

Algebra: $\mathsf{DM} = \langle A, \leq, \sim, 1, 0\rangle$ is a De Morgan algebra that satisfies analogous axioms as the one stated by the classical metatheory, however its coordinate entries have indeterminate$^{\alpha}$ membership conditions: e.g. $\exists x\exists y(\neg D^{\alpha}(x \leq y) \wedge \neg D^{\alpha}(x \not\leq y))$ is derivable. For similar reasons, $\exists x\exists y(D^{\alpha}(x \leq y) \wedge \neg D^{\alpha+1}(x \leq y))$ and perhaps even $\exists x\exists y(D^{\alpha}(x = y) \wedge \neg D^{\alpha+1}(x = y))$ are derivable.

Domain: D can be assumed to be a classical pure set in the sense of ZFC, namely the set of natural numbers; D also contains codes for sentences of the language of Field's theory.

Interpretation: The interpretation of a standard name of a sentence φ is φ. The interpretation of Tr is a mapping from D to A where every non-sentence is mapped to $\mathsf{0}$; the mapping is again indeterminate$^{\alpha}$: $\exists x\exists y(\neg D^{\alpha}(\mathsf{I}(Tr)(x) = \mathsf{I}(Tr)(y)) \wedge \neg D^{\alpha}(\mathsf{I}(Tr)(x) \neq \mathsf{I}(Tr)(y)))$. The interpretation of an n-ary arithmetical predicate is a mapping from D^n to A where the mapping is a set of classical pure set theory. All (optional) empirical and mathematical predicates are assigned sets in the standard sense, too (i.e. sets in the sense of ZFC with urelements). The arithmetical individual constants receive their intended interpretation again.

Valuation: The valuation mapping Val still yields a De Morgan-valued model, but now Val is a non-classical function. Moreover, some of the ranges of quantifiers in this model are indeterminate$^{\alpha}$ sets, too. $\mathsf{Val_S}(\varphi \rightarrow \psi) = 1 \leftrightarrow \mathsf{Val_S}(\varphi) \leq Val(\psi)$ and further constraints on the values of conditional formulas are introduced axiomatically. Of course $\mathsf{Val}(Tr(\ulcorner\varphi\urcorner)) = \mathsf{Val}(\varphi)$ is assumed as well. Furthermore,the constraints on the values of atomic arithmetical formulas are just as expected.

Prime filter: P is a prime filter of DM like the one referred to in section 3, though P has now indeterminate$^{\alpha}$ membership conditions. It is assumed axiomatically that there is no $\mathsf{x} \in \mathsf{A}$ such that $\mathsf{x}, \sim \mathsf{x} \in \mathsf{P}$. Further axioms ought to be introduced which constrain P to be a particular prime filter rather than any prime filter whatsoever.

Logical consequence: A logical consequence relation $\vdash$ is defined on the basis of the non-classical membership sign $\in$. The definition of logical consequence is still given in terms of truth preservation in all G-models and thus involves a qualification over all G-models. A *G-model* is defined to be any set theoretic model of the language of Field's theory of truth, such that (i) the domain of such a model is the set of natural numbers and (ii) the model satisfies the conditions that we stated above for the

intended non-classical model; note that the class of G-models now includes both classical and non-classical tuples of sets. A formula in the language of Field's theory of truth is called logically true if and only if it is a consequence of the empty set.

Determinate-truth$^{\alpha}$: A sentence in the language of Fields's theory of truth is defined to be determinately-true$^{\alpha}$ as follows:

$$\forall x(\mathsf{DetTr}^{\alpha}(x) \leftrightarrow D^{\alpha}(\mathsf{Val}(x) = 1))$$

(where '$\leftrightarrow$' is Field's newly defined equivalence sign).

Truth: Finally, a sentence in the language of Fields's theory of truth is defined to be true if and only if Val maps this sentence into P, i.e.:

$$\forall x(\mathsf{Tr}(x) \leftrightarrow \mathsf{Val}(x) \in \mathsf{P})$$

Discussion

Basically this metatheory is like the theory of the last section except that its logic and set theory are non-classical. One difference is that I did not define *determinate truth* simpliciter but rather *determinate-truth*$^{\alpha}$. The reason for this decision is that '$\mathsf{Val}(x) = 1$' does not have the same logical strength as it had in the classical metatheory: speaking from the viewpoint of a metametatheory of the non-classical metatheory, the sentence '$\mathsf{Val}(\varphi) = 1$' might itself have a semantic value somewhere between 1 and 0. Postulating

$$\forall x(\mathsf{DetTr}(x) \text{ iff } \mathsf{Val}(x) = 1)$$

as in the classical metatheory would leave '$\mathsf{DetTr}(\varphi)$' with precisely that value. But perhaps we want to say something that is stronger than '$\mathsf{DetTr}(\varphi)$', i.e. that φ is not just determinately true to degree 0 but rather to some ordinal degree α. For that purpose, the definition of a sequence of such indexed predicates 'DetTr^{α}' is convenient.

Why is it not so clear anymore whether Field's theory of truth is affected by a revenge problem *with respect to this non-classical metatheory*? This is because there might e.g. not be any function F such that

- $\mathsf{F}(\varphi, \psi) = 1 \leftrightarrow \mathsf{Val}(\varphi) \leq \mathsf{Val}(\psi)$
- $\mathsf{F}(\varphi, \psi) = 0 \leftrightarrow \mathsf{Val}(\varphi) \not\leq \mathsf{Val}(\psi)$
- $\mathsf{Val}(\varphi) \leq \mathsf{Val}(\psi) \vee \mathsf{Val}(\varphi) \not\leq \mathsf{Val}(\psi)$

are satisfied. Indeed, since $\leq$ is assumed to be a non-classical set, we *should* not expect the excluded middle for it. If the axioms of this non-classical metatheory are chosen in the right way, then we will perhaps be able to *prove* that there is no such function F. In this case, the fact that the language of Field's theory of truth does not include a

predicate *Imp* for which

$$\textbf{Imp1: } \mathsf{Val}(\mathit{Imp}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner)) = 1 \leftrightarrow \mathsf{Val}(\varphi) \leq \mathsf{Val}(\psi)$$
$$\textbf{Imp2: } \mathsf{Val}(\mathit{Imp}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner)) = 0 \leftrightarrow \mathsf{Val}(\varphi) \not\leq \mathsf{Val}(\psi)$$
$$\leq\textbf{-EM: } \mathsf{Val}(\varphi) \leq \mathsf{Val}(\psi) \vee \mathsf{Val}(\varphi) \not\leq \mathsf{Val}(\psi)$$

are derivable would not be considered as pointing towards the expressive 'paucity' of this language—there would be no function that could be denoted by such a predicate anyway. On the other hand, the implication sign in the language of Field's theory of truth might now turn out to be a *complete* object linguistic representation of the metalinguistic concept of partial ordering of semantic values. If we had a set theoretic principle in our metatheory that allowed us to prove that for every non-classical set there exists a classical 'digitalization' of it, then—presumably—the non-classical metatheory would be provably inconsistent. But at least from Field's standpoint it is not clear why we should have such a principle.

Accordingly, the language of Field's would not miss a classical determinate-truth predicate, but rather the non-classical metatheory could perhaps be used to show that there is nothing that such a predicate could stand for. Neither could we assume that anything is actually 'approximated' by the intended interpretations of the 'determinately-true' operators D, D^2, D^3, . . . , D^α, . . . which are defined in Field's theory of truth; the sequence would simply lack a limit and the notion of an ordinal iteration of D would perhaps neither specify a sequence of iterations uniquely nor would its extension be a classical set.

An analogous result could be shown for truth: while there would be no classical 'digitalization' of truth, the object linguistic truth predicate might yield a complete representation of metalinguistic truth in the sense that

$$\mathsf{Tr}(\ulcorner\varphi\urcorner) \leftrightarrow \mathsf{Val}(\varphi) \in \mathsf{P}$$

would be derivable by definition,

$$\mathsf{Val}(\varphi) \in \mathsf{P} \leftrightarrow \varphi'$$

might derivable set theoretically—where φ' is the metalinguistic translation of φ—and thus

$$\mathsf{Tr}(\ulcorner\varphi\urcorner) \leftrightarrow \varphi'$$

would follow (this would be possible since P is now a non-classical set). However, in order to derive such a result, it would be necessary to postulate additional constrains on P.

In such a metatheory the excluded middle should not be provable for practically any of the sets that correspond to: truth, determinate-truth$^\alpha$, the partial order of values of sentences, the concept of being a hierarchy of determinate-truth$^\alpha$ operators,

the concept of being a classical fragment of a language with truth predicate, the concept of being an iteration of D, the concept of being the ordinal number of an iteration of D, and so forth. However, if the metatheoretic axioms of this theory are chosen appropriately, it should still be possible to prove that the intended model that is described by these axioms is a (non-classical) model of Field's theory of truth.

6.5 Discussion

I have introduced two metatheories of G-solutions: one is classical, the other one is not. Now I will turn to those passages in 'Solving the Paradoxes, Escaping Revenge' which either justify my choice of axioms and definitions in these metatheories or which deviate from it significantly.

Our classical metatheory is inspired by what Field calls *G-semantics* in [2]. The models that conform to the rules of this G-semantics are nothing but classical De Morgan-valued models that satisfy formal constraints analogous to the ones that I introduced axiomatically for the intended model in section 3. The purpose of the G-semantics is to deliver a notion of logical consequence and logical truth for sentences that include the truth predicate. Since Field does not distinguish between principles that belong to G-solutions and principles that are supposed to give a metatheoretical account of G-solutions, his definition of logical consequence is stated on the same language level as his theory of truth; in contrast, I introduced the definition of logical consequence only into the metatheories of Field's theory of truth. I also added prime filters to G-semantical models as separate components; however, this is accepted as a possible option by Field (see footnote 18 on p. 17 of [2]).

There is one puzzling fact about the way Field employs his G-semantical concept of logical consequence within his theory of truth: while the latter obeys a non-classical logic, the former is defined in terms of quantification over all *classical* G-models, i.e. G-models that are pure ZFC sets. The problems of this discrepancy become much clearer once metatheories as the ones in sections 3 and 4 are introduced: in our classical metatheory, logical consequence is defined by quantifying over classical models; the intended model of Field's theory according to this metatheory is a classical model; this enables us e.g. to derive for every sentence φ in the language of Field's theory of truth: if φ is logically true, then φ is true. Accordingly, logical consequence is defined in the non-classical metatheory by quantifying over *all* models—whether models in the classical sense or models that are among the 'new' non-classical sets. The intended model of Field's theory of truth is a non-classical one. Again we derive for every sentence φ in the language of Field's theory of truth: if φ is logically true, then φ is true (given the non-classical metatheory has been spelled out successfully).

However, since Field defines consequence by reference to classical models while his intended model is non-classical, it is not clear why logical truth should entail truth at all; moreover, if it does not, what is the role of the classical G-semantical model theory in Field's theory of truth anyway? After all, the latter is not a 'supervaluation' system of partial logic and truth that is still in some sense tied to classical logic. My suggestion is thus that Field should change his definition of logical consequence to the one that was used in our non-classical metatheory of section 4 in order to avoid a blending of the classical and the non-classical metatheoretical viewpoints. But for that purpose he would first have to develop the necessary non-classical metatheory.

This non-classical metatheory is suggested by various of Field's remarks in his [2]. On p. 10 he says 'G-solutions are committed to the view that *there can be no truth-like predicate for which excluded middle can be assumed*' (his emphasis). Concerning the prospects of 'absolute' idempotent determinacy operators he states: 'I maintain that there is no good reason to think that idempotent determinacy operators . . . are intelligible' (p. 34), and 'the rationale for a G-solution would be thoroughly undermined if it could be argued that an idempotent determinacy operator is intelligible' (p. 41). A non-classical metatheory as the one in section 4 would *explain* why there can be no truth-like predicate for which the excluded middle can be assumed and also why classical determinacy operator are unintelligible—there is simply nothing there at all that could be expressed or understood! This seems to be in line with Field's view of the Liar sentence on p. 11 f. of [2] where he says: '. . . it is misleading to declare ourselves "agnostic" about the Liar sentence . . . for we don't recognize that there is a fact to be agnostic about'. On a more metaphorical level, Field is tempted to call truth a 'fuzzy prime filter' (p. 21), he speaks of 'fuzzily restricted quantifiers' (p. 55) in his theory of truth or maintains that in some cases 'the range of . . . quantification is indeterminate' (p. 58; see also p. 38). Since such 'fuzzy' prime filters and 'indeterminate' quantifier ranges are presumably simply non-classical sets, a non-classical metatheory should be used to make each of these remarks precise and indeed derivable. At the same time Field could no longer speak of sets that are indeterminate *simpliciter*—there would not be any linguistic means left to do so in such a non-classical metatheory.

In order to understand Field's theory of truth properly and in order to demonstrate its claimed revenge immunity, the introduction of a non-classical set theory along the lines sketched in section 4 is thus desirable and perhaps inevitable. Without such a metatheory, (i) the intended reference of Field's truth-theoretic vocabulary remains unclear: e.g. what kind of non-classical set is the extension of *Tr*? (ii) Field's remarks from above remain on a merely metaphorical level, whereas a non-classical set theory would *explain* properties such as non-intelligibility, non-factivity, and fuzziness that Field refers to; (iii) as seen above, logical consequence remains defined inaccurately

by quantification over *classical* G-models only, which has the effect that logical truth does not necessarily imply truth.

If the systematic development of such a theory is found to face insurmountable obstacles, Field's whole project is cast into doubt. Field's naive theory of properties in [5] can be regarded as a first attempt at such a metatheory, however, due to its lack of the extensionality axiom, as a preliminary one. Additionally, [5] is almost completely silent about what the intended identity relation for properties is like, although we are reminded by Evans' [1] classical paper that claims of vague or 'fuzzy' identity of objects might contradict time-honored principles such as Leibniz' Law or λ-conversion. As Field himself points out when he discusses the restricted version of separation on p. 14 of [2], the step from a non-classical theory of properties to a non-classical theory of sets is non-trivial: 'Requiring excluded middle as an assumption of separation seems reasonable: otherwise, we would license sets for which membership in the set depends on whether the Liar sentence is true; given extensionality, this would lead at the very least to indeterminate identity claims between sets, and it isn't at all clear that paradox could be avoided even allowing that'. What Field does not acknowledge in this quotation is that precisely such a theory has to be developed in order to 'ground' his theory of truth both formally and philosophically. If it turned out that Field's theory of truth does not allow for such an extensional interpretation, then the ultimate price of accepting the theory would have finally become transparent, as Field would find himself in the position to presuppose—*nolens volens*—a non-extensional semantics in order to adhere to his choice of type-free theory of truth. *Vengeance is mine, I will repay, saith the Metatheory.*

References

[1] Evans, G. (1978). 'Can there be vague objects', ***Analysis*** 38, 208

[2] Field, H. 'Solving the paradoxes, escaping revenge', this volume

[3] ——— (2003). 'A revenge-immune solution to the semantic paradoxes', *Journal of Philosophical Logic* 32, 139–77

[4] ——— (2003). 'The semantic paradoxes and the paradoxes of vagueness'. *Liars and Heaps*. In JC Beall (ed.), Oxford University Press, pp. 262–311

[5] ——— (2004). 'The consistency of the naive theory of properties', *Philosophical Quarterly* 54, 78–104

I would like to thank Hartry Field, Philip Welch, Sol Feferman, Graham Priest, Leon Horsten, and George Bealer for very helpful discussions on this chapter.

[6] ____ (2005).'Variations on a theme by yablo' *Deflationism and Paradox* In JC Beall and B. Armour-Garb (eds.), Oxford University Press

[7] Hajek, P. Paris, J., and Sheperdson, J. (2000). 'The liar paradox and fuzzy logic', ***The Journal of Symbolic Logic*** 65, 339–46

[8] Kripke, S. (1975). 'An outline of a theory of truth', *Journal of Philosophy* 72, 690–716.

[9] Leitgeb, H. (1999). 'Truth and the liar in de morgan-valued models', ***Notre Dame Journal of Formal Logic*** 40, 496–514

[10] Priest, G. (2005). 'Spiking the field artillery'. In JC Beall and B. Armour-Garb, eds., *Deflationism and Paradox* Oxford University Press

[11] Welch, P. 'Ultimate truth vis a vis stable truth', *submitted*

[12] Yablo, S. D. (2003). 'New grounds for naive truth theory'. In JC Beall (ed.), *Liars and Heaps* Oxford University Press, pp. 312–30

7

Reducing Revenge to Discomfort

Tim Maudlin

My understanding of the Revenge problem is simple and prosaic. One begins with a stock of semantic predicates, for example 'true' and 'false', and uses some of them to construct a problematic sentence such as '*F* is false', where *F* is stipulated to refer to that very sentence. One then considers the various hypotheses that attribute the different semantic predicates to the sentence: '*F* is true' and '*F* is false' in this instance. Each of these hypotheses is shown, by informal reasoning, to lead to an impossibility. Suppose we grant that no sentence is both true and false. Then since '*F* is true' implies, by disquotation, *F* (i.e. '*F* is false'), the hypothesis that *F* is true implies that *F* is both true and false, an impossibility. So we cannot maintain that *F* is true. But similarly, the hypothesis that *F* is false leads to the same impossibility, by anti-disquotation, as it were. (The principles of disquotation and anti-disquotation correspond to the Downward and Upward T-Inferences in my *Truth and Paradox* (2004), and are also called T-Out and T-In or T-Elim and T-Intro.) We seem to have only two choices: decide that some sentences *can* be both true and false, so the conclusion is not impossible after all, or conclude that the sentence is neither true nor false, since each alternative leads to an impossibility. The first route is dialethic, and the second we might call standard. A standard reaction must therefore be to introduce some *new* semantic category, a category of the neither-true-nor-false. Let us call this new category, for the moment, simply *Other*. The standard response to this problem, then, is to pronounce the problematic sentence neither true nor false, but Other.

Having started with a stock of two semantic predicates (in this example), we are forced to increase the stock. And now the elements of Revenge are in place. Using the new semantic predicate, we are in a position to construct a new problematic sentence, e.g. '*F-or-O* is either false or Other', where *F-or-O* is stipulated to denote that very

sentence. Once again, we assume that the semantic categories exclude each other, so no sentence is both true and Other, or both false and Other, or both true and false. Now we repeat the argument: the hypothesis that *F-or-O* is true implies that it is either both true and Other or both true and false, each of which is impossible. The hypothesis that *F-or-O* is false implies it is either false or Other, and hence is true. The hypothesis that *F-or-O* is Other similarly implies it is either false or Other, and hence true. Since each alternative leads to an impossibility, we must reject all three. *F-or-O* is neither true nor false nor Other. So it must be something else. We need a new semantic category, call it Weird, which then allows for the construction of the new problematic sentence '*F-or-O-or-W* is either false or Other or Weird', and off we go on another cycle. In each turn of the crank we add a new semantic category, and then add that new category as an alternative in a new problematic sentence.

Approaching the situation in this way, we seem to reach a conclusion: for any given set of possible semantic values, if they are mutually exclusive, then they are not jointly exhaustive. Supposing that no sentence can have more than one value implies that at least one sentence has a semantic value not on the list. The list must grow, without end.

If we endorse the puzzle as presented in this form, there appear to be only two choices. One is to accept that there can be no set of semantic values that is jointly exhaustive: the hierarchy of semantic values (and hence a hierarchy of languages that contain predicates for semantic values) simply has no end. For every language, there is a semantically richer metalanguage. Or, on the other hand, one can reject the claim that all the semantic values are mutually exclusive. If we accept that certain sentences can have more than one semantic value, then some of our argumentation breaks down. We posit that the problematic sentence *F* is true and derive that it is false: very well, perhaps it is both true and false! If this is not an impossibility, then we have no *reductio* of the original supposition. Our choice, it seems, is between an unending progression of semantic values or dialethism.

But even these rather unpalatable options do not provide relief from Revenge. So long as one presents the problem in this way, it seems at least coherent to imagine accepting either that any list of semantic values is incomplete or that a single sentence can have more than one value. If one thinks very abstractly of a semantic theory as just assigning *letters* (**T, F, O, W**, or whatever) to sentences, then one can imagine adopting a scheme that either assigns several letters to a single sentence or that fails to assign any from the set at all. But all of these options seem to be foreclosed if the problematic sentence does not *attribute* to itself semantic values other than truth, but rather *denies* of itself that it is true. Consider the following passage by Tyler Burge:

> Consistent, nonbivalent logics with a univocal truth predicate are certainly constructible. But no such logic, insofar as it assumes a truth predicate with a constant extension, has given a plausible account of the semantical paradoxes. This is because of a family of problems that have

come to be known as the 'Strengthened Liar'. *The Strengthened Liar* (perhaps better called 'The Persistent Liar') is really the original Liar reiterated for the sake of those who seek to undercut paradox primarily by appeal to a distinction between falsehood and some other kind of truth failure. Failure to resolve the Strengthened Liar is not a difficulty of detail or a mere drawback in a solution. It is a failure to account for the basic phenomenon. Any approach that suppresses the liar-like reasoning in one guise or terminology only to have it emerge in another must be seen as not casting its net wide enough to capture the protean phenomenon of semantical paradox.

The Strengthened Liar in its simplest form is this. If we analyze

$$(\beta)(\beta) \text{ is not true}$$

as being neither true nor false, then it intuitively follows that the sentence displayed is not true. But the sentence displayed is (β). So it seems to follow that (β) is not true after all. We have now apparently asserted what we earlier claimed was neither true nor false. Moreover, the assertion that (β) is not true would seem to commit us to asserting '(β) is not true' is true, contrary to our original analysis. It is important to see that this informal reasoning is entirely intuitive.

Burge (1979), pp. 172–3

The Strengthened Liar does not make use of any semantic predicate beside truth, as *F* and *F-or-O* did. The Strengthened Liar does not *assert* of itself some semantical value other than truth, it rather merely *denies* that it is true. So the relevant semantic categories are not truth and falsity, but rather truth and absence of truth. And these two categories *do* seem to be clearly mutually exclusive and jointly exhaustive: every sentence, indeed every object, is either among the things that are true or it is not. Both conditions can't obtain: if something is true, then perhaps it could possibly also be false, or also be Other, or also be Weird, but it cannot fail to be true. And if it fails to be true, then it is not true: the categories are jointly exhaustive. The whole apparatus of adding new semantic values never arises, and the dialethic alternative is unavailable. It may merely be hard to comprehend how a sentence could be both true and false, but even accomplishing that feat would do nothing to make comprehensible how a sentence could both be true and fail to be true.

A natural reaction at this point is to shift focus a bit. If the paradoxes ensue simply from the supposition that any given sentence is either true or not true, then one considers how even that claim could be denied. Perhaps 'true' is a vague predicate, with no determinate boundary separating sentences that satisfy it from those that don't; perhaps it is somehow unstable, with a constantly shifting extension. Solutions along these lines have been pursued. So, in very schematic form, we have already generated a series of approaches to the Liar, involving hierarchies, dialethism, and vagueness. All of these strategies agree on the following: there is not one, single, univocal truth predicate with an exact, fixed extension such that no true sentence can have a semantic value other than true.

According to the account of truth I have developed in *Truth and Paradox*, there is a single, univocal truth predicate with an exact, fixed extension such that no true sentence can have a semantic value other than true. Or, slightly more exactly, if there are cases of indeterminate truth-value that derive from vagueness, they have nothing to do with the classical semantic paradoxes. The solution of the paradoxes does require the introduction of a novel semantic value, but just one. The Liar paradox can be completely and adequately addressed in a language with three semantic values: true, false, and Ungrounded. Each of these semantic values has a perfectly determinate extension, and the three semantic values are mutually exclusive and jointly exhaustive.

All of the alternative approaches agree, one way or another, that the reason we get into trouble with paradoxical reasoning is that there is something screwy with the concept of *truth*. Maybe there are an infinitude of truth predicates, or the extension of the predicate is hazy or fluctuating or context-dependent, or maybe, contrary to our initial intuitions, truth does not preclude falsity. In contrast, I think that it is rather evident that the reason we get into trouble with paradoxical reasoning is that there is something screwy with the *reasoning*. As Burge notes above: 'It is important to see that this informal reasoning is entirely intuitive'. That is, the reasoning is reasoning that we are inclined to accept, without further argument, as valid. But since the reasoning leads us into unacceptable conclusions, our intuitive acceptance of it must be incorrect. We must, in fact, be making mistakes in our reasoning, albeit mistakes which are not immediately evident.

The unreliability of intuitively correct reasoning is starkly illustrated by the argument form known as Löb's paradox or Curry's paradox. This argument form apparently allows one to prove any sentence at all from no premises whatsoever. Suppose, for example, I wish to prove by purely logical considerations that the moon is made of green cheese. All I need to do is introduce a sentence, which we'll call 'Sam'. What Sam says is this: If Sam is true, then the moon is made of green cheese. (The use of the proper name 'Sam' for the sentence is incidental to the argument—one could, for example, replace the proper name with a definite description that, as a matter of contingent fact, denotes that very sentence.) The entirely intuitive reasoning runs as follows. Suppose, for the sake of argument, that Sam is true. Then we have both that Sam is true (by hypothesis) and if Sam is true, then the moon is made of green cheese (since that's what Sam says). From these two claims together, we can derive, by modus ponens, that the moon is made of green cheese. So from the supposition that Sam is true, we can derive (by entirely intuitive reasoning) that the moon is made of green cheese. That is, the proposition that the moon is made of green cheese follows logically from the proposition that Sam is true. But whatever follows logically from a true proposition is true. So if Sam *is* true, the moon is made of green cheese. This conclusion, derived by entirely intuitive reasoning, is, of course, Sam itself! So if the

intuitive reasoning is to be trusted, Sam is true. But it follows from that (as we have seen), that the moon is made of green cheese. So it is a conclusion of entirely intuitive reasoning, from no premises at all, that the moon is made of green cheese. QED.

We know that our *reasoning* is unreliable here. The question is: what is wrong with it, and how do we fix it? Burge seems content to leave the reasoning at the *informal* level, but there is no impediment to formalizing the argument. All the argument uses are standard logical inferences and a pair of inferences concerning truth: from the claim that a sentence is true we can infer the sentence itself, and from any sentence we can infer that the sentence is true. These inferences are the *Downward* and *Upward T-Inference* respectively. These inferences are intuitively valid, i.e. truth-preserving. For if a sentence is true, then surely the sentence that *says* it is true is also true. And if it is true to say a sentence is true, then that sentence is true. All that we need to accept about truth in order to get the paradoxes going is the validity of these inferences *since these inferences, together with standard logical principles, are all that is employed in the paradoxical reasoning*. The Green Cheese argument, for example, can be formalized as follows:

$T(Sam)$	Hypothesis
$T(Sam)$	Reiteration
$T(Sam) \supset G$	Downward T-Inference
G	$\supset$ Elimination
$T(Sam) \supset G$	$\supset$ Introduction
$T(Sam)$	Upward T-Inference
G	$\supset$ Elimination

'G' stands for the claim that the moon is made of green cheese, although it could evidently be anything you like. *T* is the truth predicate. All one needs for the derivation, beside the standard logical inferences and the *T*-Inferences, is the information that '*Sam*' denotes the sentence '$T(Sam) \supset G$'.

We now have a formalized argument, all of whose inferences are intuitively acceptable, proving that the moon is made of green cheese. This is serious trouble, trouble that must be dealt with by any account of the paradoxes. But note that some approaches don't seem, at first glance, to make any contact with the problem at all. How, for example, would it help matters to go dialethic and grant that some sentences are both true and false, or that some contradictions are true? Our argument does not employ any contradictions, and the problem is that it appears to establish a conclusion that is completely, unproblematically, false. If an account of truth cannot return the result that 'The moon is made of green cheese' is false *and not true*, then it is of no use at all.

Similarly, having the extension of the truth predicate be vague or fluctuating doesn't seem to help. Let the predicate *T* stand for 'True in a completely determinate,

unproblematic, unfluctuating way, the way 'The moon is not made of green cheese' is true'. The sentence Sam may not be true in that way, but that is neither here nor there: all we are doing is reasoning from the *supposition* that it is. The reasoning appears to establish that the moon is made of green cheese. (One would need the additional claim that logical truths that can be proven by 5-step arguments are true in a completely determinate, unproblematic, unfluctuating way.)

Focusing on this *argument* is a healthy corrective to Burge's approach. Burge produces an intuitively acceptable argument to an unacceptable conclusion, then goes on to worry about the concept of truth, rather than the flaws in the argument. But if we know that the argument is flawed, then we obviously can't draw *any* conclusion from it except that it contains an error. Once we have formalized the argument, we can then search for the error. *And the error does not occur in the use of the T-Inferences*. The error occurs in the *classical* inferences, and does not impugn our intuitions about truth.

To make a long story short, some classical inferences tacitly presuppose, for their validity, that the language is bivalent, i.e. that every sentence is either true or false, and hence if a sentence is not true, it is false. These inferences are particularly easy to spot in a natural deduction system: they are exactly the inferences that use subderivations and thereby allow the proof of theorems. The culprit in the argument above is the rule for $\supset$ Introduction. This rule uses a subderivation that can begin with any arbitrary hypothesis H. If one is able to derive any other sentence, C, then one can dismiss the subderivation and write $H \supset C$. By what rights should one be able to do this?

The usual *justification* of the rule is this. Suppose the inferences used in the subderivation are all valid. Then if H happens to be true, C will be true, and so $H \supset C$ will be true. But if H is not true, then (HERE COMES BIVALENCE!) H is false, and any conditional with a false antecedent is true. So either way, $H \supset C$ is true.

But what if the language is not bivalent? Then H can fail to be true without thereby being false, and the second part of the justification fails. Suppose that H, besides being true or false, could also be Ungrounded. Suppose further that a conditional with an Ungrounded antecedent and consequent is also Ungrounded. Then the fact that one can validly derive C from H does not imply that $H \supset C$ is true. The 'entirely intuitive' inferential principle of $\supset$ Introduction can be invalid. (Note: by use of $\supset$ Introduction and simple reiteration, one can prove $H \supset H$ as a theorem, for any H. If $H \supset H$ is Ungrounded when H is Ungrounded, this conclusion need not be true, so $\supset$ Introduction is not valid.)

In *Truth and Paradox*, I argue that, *given the meaning of ungroundedness*, any logically complex sentence all of whose atomic constituents are Ungrounded must itself be Ungrounded. The basic picture is that logically complex sentences—sentences that contain logical particles—have their truth-values determined by the truth-values of other sentences. For example, the truth-value of a conjunction is a function of the truth-value of the conjuncts, the truth-value of a negation a function of the truth-value of the sentence

negated, and so on. This allows us to draw a directed graph of a language, with arrows running from the immediate semantic constituents of every logically complex sentence to that sentence (e.g. from the conjuncts to the conjunct, from the negated sentence to the negation). The key is to treat the *truth predicate* as a logical particle in just this way, where the immediate semantic constituent of a sentence of the form $T(\alpha)$ is the sentence denoted by α. That is, although $T(\alpha)$ is grammatically atomic, it is logically complex. The truth predicate, like conjunction, disjunction, and negation, is associated with a truth function, viz. the identity map from the truth-value of the sentence denoted by α to the truth-value of $T(\alpha)$. One effect of treating the truth predicate as a logical particle in this way is to allow the directed graph of a language to contain cycles. But more generally, the picture is this: the graph of a language will typically have a boundary: a set of logically atomic sentences whose truth-values are determined not by the truth-values of other sentences but by the world. The truth-values of the logically complex sentences are then determined from the truth-values of the boundary sentences by means of truth functions. But some sentences, like the Liar and the Truthteller, will never inherit a truth-value from the boundary of the graph because they are not connected, via the graph, to the boundary. These sentences are ungrounded (as are others that may be connected to the boundary, but not in the right way to have their truth-values determined, by means of truth functions, from the truth-values of the boundary sentences).

Given this account of ungroundedness, it follows that any logically complex sentence all of whose immediate semantic constituents are Ungrounded will itself be Ungrounded. So the truth table of any logical particle must map Ungrounded input into Ungrounded output: a conjunction or disjunction of Ungrounded sentences must be Ungrounded. And similarly for the conditional. So if H is Ungrounded, so is $H \supset H$.

Similar argumentation shows that in a trivalent language the rule of $\sim$ Introduction will not be valid. That is the rule used in standard reasoning about the Liar. Consider, in particular, the Strengthened Liar. Let λ denote the sentence $\sim T(\lambda)$. Then we can apparently *prove* contradictory sentences by the following simple argument:

$T(\lambda)$	Hypothesis
$T(\lambda)$	Reiteration
$\sim T(\lambda)$	Downward T-Inference
$\sim T(\lambda)$	$\sim$Introduction
$T(\lambda)$	Upward T-Inference

If the Downward and Upward T-Inferences are valid, then the only culprit can be $\sim$ Introduction. And again, what is the justification for this rule? Grant that no sentence and its negation can both be true. Then if one reasons validly from a true hypothesis,

one cannot possibly derive both a sentence and its negation. So if one *can* validly derive both a sentence and its negation from a hypothesis, the hypothesis cannot be true. *In a bivalent language, this implies that the hypothesis is false, and hence its negation is true*. And this would justify the rule of $\sim$ Introduction.

But again, suppose the language is trivalent, with the semantic value Ungrounded, which has the property that the negation of any Ungrounded sentence is also Ungrounded. Then the justification fails: one might be able to validly derive contradictory sentences from an Ungrounded hypothesis without the negation of the hypothesis being true. $\sim$ Introduction is intuitively acceptable because almost all of our everyday thought and argument is in the bivalent sector of the language, where use of the rule will generate only truths. But when the topic of the Liar comes up and we wander into the non-bivalent sector the invalidity of the argument form becomes important.

One advantage of focusing on the *validity* of the reasoning about the Liar is this: it sidesteps the problem of Revenge. What we saw initially was that Revenge is most straightforward when one *accepts the reasoning and tries to accommodate the conclusion by adding a new semantic value*. One reasons that the Liar, in the form 'This sentence is false', is not true and is not false, since each alternative leads to a contradiction. One accepts the conclusion, and adds a new alternative. But Revenge then shows that one's concessions have not ended: a revised Liar sentence will force acceptance of yet another new semantic value, and another, without end. But if one *rejects the conclusion* of the reasoning because one *rejects the reasoning*, then Revenge has no obvious foothold: simply *repeating* the original reasoning with new semantic values will have no force, since it is repeating an invalid argument.

In order to get a foothold again, the revenger must now try to *revalidate the reasoning*. This, however, is a tricky business. We saw above that the Löb's paradox reasoning relies on $\supset$ Introduction, and that rule, in turn, is normally justified by appeal to bivalence. But now the revenger can try to reintroduce the problem not by invoking a new truth-value but rather by invoking *new logical operators*. Suppose, for example, that there were any two-place operator $\Rightarrow$ such that (1) whenever $A \Rightarrow B$ is true and A is true, B is true, (so modus ponens is valid for $\Rightarrow$); (2) whenever both A and B are true, $A \Rightarrow B$ is true, and (3) whenever A is anything other than true, $A \Rightarrow B$ is true. Then the analog of the reasoning given above cannot be diagnosed the same way: $\Rightarrow$ Introduction and $\Rightarrow$ Elimination would both be valid. Or suppose we require condition (1), condition (2'): whenever B can be validly derived from A, $A \Rightarrow B$ is true. Then again, the reasoning above cannot be blocked. So it is a burden of this approach to the Liar to argue that *no logical operator can satisfy (1), (2), and (3) or both (1) and (2')*. The reason for this has been given above.

Ungrounded sentences are sentences whose truth-value is not determined by the truth functions associated with the logical particles even when the truth-values of all the atomic sentences of a language have been settled. We consider the truth

predicate to be a logical operator, so no sentence containing the truth predicate is logically atomic. It follows that for any truth-functional operator *if all the arguments of the operator are Ungrounded, so is the sentence whose main connective is that operator*. If A and B are both ungrounded, so is $A \Rightarrow B$, provided only that $\Rightarrow$ is a logical (truth-functional) operator. This blocks the revenger's attempt to revalidate the inference.

The resulting semantic theory is essentially Kripke's theory using the minimal fixed point. If one accepts that theory and uses it to determine when syntactically specifiable inferences are valid, one finds that the T-Inferences are valid while various classical inferences are not. The Liar reasoning is defused, and with it all the Revenge reasonings as well.

Kripke himself was not satisfied with this theory. According to the theory, the Liar sentence is not true: it is Ungrounded. Hence the sentence that says the Liar is not true—i.e. the Liar sentence itself!—is ungrounded. So if we want to *assert* that the Liar sentence is not true, we want to assert an ungrounded sentence. And Kripke did not like this consequence: he only wanted to assert sentences that are true. This led him, in the end, to suggest 'closing off' the truth predicate, i.e. introducing a *new* predicate True* such that the Liar sentence, while not true, would be True*. This is an entirely *new* form of Revenge, and we need to consider it carefully.

Here is what Kripke has to say about the shortcomings of the minimal fixed-point theory:

> [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. If we think of the minimal fixed point, say under the Kleene valuation, as giving a model of natural language, then the sense in which we can say, in natural language, that a Liar sentence is not true must be thought of as associated with some later stage in the development of natural language, one in which speakers reflect on the generation process leading to the minimal fixed point. It is not itself part of that process. The necessity to ascend to a metalanguage may be one of the weaknesses of the present theory. The ghost of the Tarski hierarchy is still with us . . .
>
> Since the object language obtained by closing off $T(x)$ is a classical language with every predicate totally defined, it is possible to define a truth predicate for that language in the usual Tarskian manner. This predicate [which I above call True*—T.M.] will *not* coincide in extension with the predicate $T(x)$ of the object language, and it is certainly reasonable to suppose that it is really the metalanguage predicate that expresses the 'genuine' concept of truth for the closed-off object language; the $T(x)$ of the closed-off language defines truth for the fixed point *before* it was closed off. So we still cannot avoid the need for a metalanguage.
>
> Kripke (1975) pp. 714–15

Although Kripke does not mention it, Revenge lurks just around the corner. For if True*(x), the metalanguage predicate, expresses the 'genuine' concept of truth, then

'This sentence is not True*', rather than 'This sentence is not true', expresses the 'genuine' Liar sentence, and all of our analysis of 'This sentence is not true' has been for nought. We have not just 'the ghost of the Tarski hierarchy', but the full-fledged complete hierarchy, forced on us by Revenge. Consideration of 'This sentence is not true' leads us to introduce the metalanguage predicate True*, which affords the construction of 'This sentence is not True*', and so on.

Let's consider very carefully what led Kripke to this unpalatable position. 'Liar sentences are not true in the object language' Kripke says, '. . . but we are precluded from saying this in the object language by our interpretation of negation and the truth predicate'. In what sense *precluded*? We can obviously *say*, in the object language, that the Liar sentence is not true: *this is exactly what the Liar sentence itself says*! The object language, in this case, contains a truth predicate, and contains negation, and contains individual terms and descriptions that denote the Liar sentence. These afford all the resources one needs to *say* that the Liar is not true, by means of the Liar itself.

So the problem that Kripke has cannot be that one can't say that the Liar is not true. Rather, it must be that one cannot *truly* say that the Liar is not true. And this is perfectly correct: since the truth-value of the Liar turns out to be Ungrounded, one cannot *assert* the Liar without thereby asserting an Ungrounded sentence, rather than a true one. Kripke's problem is not that the claim is inexpressible in the object language, but that the claim, so expressed, is not *true*.

So it must be that Kripke thinks that in order to be *entitled* to make a claim, the claim must, in some substantial sense, be true. He can't get the sentence 'The Liar is not true' itself to come out true, of course, but by closing off the truth predicate he can get it to come out True*. And since the whole idea is to *justify the assertion that the Liar is not true*, i.e. justify the assertion of the Liar itself, it then has to turn out that, for purposes of making assertions, what's important is not being true but rather being True*. This presumably explains the sense in which the 'metalanguage predicate . . . expresses the ''genuine'' concept of truth for the closed-off object language'. But now we are down the rabbit hole with no return: the 'genuine' form of the Liar ought to employ the 'genuine' truth predicate. And if justifiable assertability goes with Truth* rather than with truth, we are left wondering why we were worried about the original Liar sentence at all.

I think that Kripke has gone off the tracks here. There is only one truth predicate, the genuine and important one, and it is there in the 'object' language. And the semantic theory shows, quite clearly, that the Liar sentence is not true. And we can *express* this perfectly by asserting the Liar sentence itself. The only question left is the *justification* of this assertion. It evidently cannot be justified by appeal to the truth of what is asserted: can it be justified in some other way?

Exactly this problem can be found in Burge's presentation as well. By means of a chain of reasoning—let's not worry about its validity now—Burge concludes 'We

have now apparently asserted what we earlier claimed was neither true nor false. Moreover, the assertion that (β) is not true would seem to commit us to asserting "(β) is not true" is true, contrary to our original analysis.' I agree with Burge that at the end of the argument we assert what we earlier claimed was neither true nor false—viz. the Liar sentence. Furthermore, I claim we can *correctly* assert it, even though it is neither true nor false. And that once we understand properly the *grounds* on which the assertion is correct, we see that it *does not* 'commit us to asserting "(β) is not true" is true'.

The question before us—the only barrier between us and a resolution to the Liar paradox—is a *normative* question: under what circumstances is one *entitled* to assert a sentence. If the only circumstance is when the sentence is true, then we have the problem Kripke falls into. But is that really the only reasonable option?

We should first note that there is no question here of trying to justify the assertion of a *false* sentence. The Liar sentence is Ungrounded—neither true nor false—so the assertion of it would not misrepresent facts in the way assertion of a false sentence does. Since most everyday discourse is bivalent, in everyday contexts lack of truth is equivalent to falsehood. Restricting oneself to asserting only truths is, in this context, just the same as avoiding the assertion of falsehood. But if there are sentences that are neither true nor false, everyday normative attitudes towards them are likely to be non-existent. We are likely to be as much at sea about how to properly handle *assertion* of such sentences are we are evidently at sea about how to *reason* about them.

What we need are normative rules governing the assertion of sentences, rules that go beyond 'Assert the true and do not assert the false.' What we want, I argue, are rules that permit the assertion of some Ungrounded sentences, e.g. the Liar sentence, and forbid the assertion of other Ungrounded sentences, e.g. the negation of the Liar or the claim that the Liar is true. And such a set of rules is not hard to come by: the most obvious set corresponds, in a straightforward sense, to what Kripke tries to achieve by closing off the truth predicate. These rules can be articulated, and it is unproblematically true, for example, that assertion of the Liar sentence *is permitted by these rules*, while the claim that the Liar sentence is true *is forbidden by these rules*. Even though, according to Burge, the assertion of the Liar sentence 'would seem to commit us to asserting' that the Liar sentence is true, according to these rules it does not do so. So these rules of permissible assertion, at any rate, avoid Burge's presentation of the Strengthened Liar.

Of course, these rules are just one set among an infinitude of possibilities. 'Only speak the truth' is another, albeit one that would leave us rather tongue-tied in the presence of the Liar: one could neither say it is true, nor not true, nor false, nor not false, nor neither true nor false, etc. (The rules that I endorse do permit us to say of the Liar that it is neither true nor false. Indeed, 'neither true nor false' is essentially equivalent to 'Ungrounded', so we don't need to introduce 'Ungrounded' into our

language if it already has a truth and falsity predicate. We can permissibly—but not truly—say that the Liar is Ungrounded.) In face of all these various sets of normative rules, our problem then is simply to *choose between them*. In *Truth and Paradox* I approach this problem by articulating an Ideal, a set of properties that we would like the normative rules to have. It is demonstrable that no set can have *all* the Ideal properties, but there are sets that have many. Among them is the particular set I advocate.

And, lo and behold, at the end of the day, all of this leads to yet another form of Revenge, but Revenge is a somewhat mutated form. Let me finish by discussing the problem.

Consider any set of normative rules for the assertion and denial of sentences, call the set *R*. Among the properties in the Ideal, properties we would like *R* to have, are these:

(1) *R* should permit the assertion of all true sentences.

(2) *R* should forbid the assertion of all false sentences.

(3) *R* should be pragmatically coherent, i.e. it should not both permit and forbid the assertion of any sentence.

(4) *R* should be complete: it should, in any particular circumstance, render a judgement about whether any sentence ought or ought not to be asserted.

I take it that it is uncontroversial that these *are* elements in the Ideal: we would like the normative rules to satisfy these demands. Failure to meet any one of them is a failure to achieve the Ideal.

But it is easy to see that *any* set of normative rules must fall short of the Ideal one way or another. For any particular set of rules *R*, consider the sentence

According to the set of rules *R*, one is not permitted to assert this sentence.

What have the rules in *R* to say of this sentence?

Since we have in no way specified how *R* deals with this sentence, we cannot draw any conclusions, but no matter how the rules are written, they are bound to fall short of the Ideal. For if the rules permit the assertion of the sentence, they permit the assertion of a falsehood. Thus they forbid its assertion, they forbid the assertion of a truth. If they both permit and forbid the assertion, they are pragmatically incoherent. If they render no judgement at all, then they are incomplete. Similarly if they are vague in a way that makes their application uncertain. Each of these failures is a flaw relative to what we would like the rules to achieve.

Failure to achieve an Ideal, when achievement is logically impossible, is not a paradox. An analogy: suppose I set out to write a program for a computer to play chess. One might ask about the Ideal program, the program that would satisfy one's wildest dreams. That would clearly be a program guaranteed to win every game it

plays, whether as White or Black. Certainly, one *wants* the computer to win every game it plays, and designs the program with that as the goal. But it is easy enough to see that the Ideal program is a logical impossibility: playing against itself, it would have to win both as White and as Black. This is not a paradox—nor does it undercut the Ideal as a kind of standard. It does mean, though, that the Ideal will never be achieved: one has to settle, no matter what one does, for something less. What we have in the end is simple disappointment.

So Revenge gets its flesh, but it is an ounce rather than a pound. Rules for permissible assertion of sentences can be found that deal with the original Liar in a perfectly adequate way, requiring us only to recognize that, since some sentences that are not true are permissible to assert, sometimes it is permissible to assert a sentence but not permissible to assert that the sentence is true. This allows us to solve both Burge's and Kripke's problems. But the solution demands that we recognize a new status, beside truth or falsity or Ungroundedness, that a sentence may have: viz. being permitted by a particular set of rules. And once we introduce a predicate that allows us to express this status in our language, we can construct a Liar-like sentence that denies of itself that it has this status. This Liar-like sentence does cause some trouble.

But the trouble only amounts to showing that our rules cannot completely satisfy the Ideal, and this is trouble that we can learn to live with. Accepting 'intuitive reasoning' that allows us to prove any sentence whatever is not an option: Löb's paradox and seemingly valid reasoning about Liar sentences that leads to contradictions must be diagnosed and blocked. That job can be done. We are left only with a few pathological sentences that, by their construction, cannot be dealt with by rules for permissible assertion in an ideally satisfactory way. Since these sentences are of no other interest, it does us no harm to pay them little heed. As the sage saith: You can't always get what you want, but if you try sometimes you might find you get what you need.

References

Burge, T. (1979). 'Semantical paradox'. *Journal of Philosophy* 76, 169–98
Kripke, S. (1975). 'Outline of the theory of truth'. *Journal of Philosophy* 72, 690–716
Maudlin, T. (2004). *Truth and Paradox: Solving the Riddles*. Oxford: Oxford University Press

8

Understanding the Liar

Douglas Patterson

It seems natural to assume that understanding a language like English is knowing what its sentences mean and, since generally when someone knows something it can be said what it is, to assume that understanding a language is bearing some relation to a true theory that states what its sentences mean. Sentences such as 'this sentence is false' and expressions such as 'does not apply to itself' raise problems as to how any such theory could be true, since they induce the derivation of contradictions from claims about them that it would seem reasonable or even unavoidable to include in an adequate semantic theory for a language that contains them. Indeed, Tarski flatly asserted that the set of true sentences could not be defined for any language containing such expressions (1944, 347–9; 1983, 165, 247 ff.), a view which, if correct, undercuts the possibility of any account of meaning on which meaning is understood at least in part in terms of the conditions under which a sentence would be in the set of true sentences of a language. Accepting this, however, goes against the assumption that understanding a language like English is some relation to a true theory of this sort, and the traditional response has been to blame Tarski's views on the simplicity of his logic, in particular on the classical partition of sentences into the true and the false. Some additional category, it is thought, such as 'meaningless' or 'does not express a proposition', is required to handle such sentences.

The response in turn gives rise to what is known as 'the revenge problem', in which each approach of this sort, applied to natural languages, runs up against further problematic expressions that render the approach itself again inconsistent. Calling 'this sentence is false' meaningless might look workable at first. Calling problematic sentences 'meaningless' won't do, however, for 'this sentence is either false or meaningless', which the approach must ban from the language to which it

applies on pain of further contradiction or untruth in its own implications.[1] I believe that the revenge problem is insoluble and that the search for an approach of this sort rests on the assumption with which I began. I suggest a different tactic: take understanding to be an attitude that like belief and unlike knowledge may be held falsely and allow that understanding a natural language is actually bearing such a non-factive cognitive relation to a semantic theory shown by the paradoxes simply to be false. I will argue that on this approach we can do everything we ever needed to do in natural language semantics while avoiding the revenge problem entirely.

What I will be offering is a variant of 'inconsistency' views of the semantic paradoxes. Supporters of such views are Charles Chihara (1979, 1984), Matti Eklund (2002, 2005), Jody Azzouni (2003), and Tarski himself.[2] I have, therefore, some heavy lifting to do: such views are widely believed to be not just false, but ridiculous; as Azzouni notes, 'despite Tarski's status, the *scorn* (over the years) heaped on this particular view of his is impressive' (2003, 337), and he is able to bring out familiar quotes from Burge, Katz, Putnam, and Parsons accusing the view of resting on obvious confusion or implying obvious absurdity. Indeed, it is rather common to believe that the fact that Tarski spoke, wrote, and was understood in natural language is sufficient refutation of his view. I believe, however, that the main reason that the inconsistency view has seemed unacceptable is that it conflicts in an obvious way with the claim that understanding is a kind of knowledge: if you know what the sentences of English mean, and what you know is somehow inconsistent, then you know something false, which really is absurd. My main point will be that if understanding a language can be a relation to a false semantic theory, then the inconsistency view, in the form of the claim that fully competent speakers of certain languages understand them in accord with a theory that could be true only if contradictions, and indeed everything, were, survives the traditional objections and in fact emerges as the best response to the paradoxes.[3] Along the way, of course, I will explain how communication is possible in a language of this sort. Note in particular that my view will be not that English *is* somehow 'inconsistent', but that competent speakers take it to have features that it could have only if it were.[4]

[1] See §8.4 for more here. Scharp (this volume) provides a very articulate discussion.

[2] Heck (2004) also seems to endorse the view in its concluding paragraph; I will comment on this in a note below.

[3] Of late views on which contradictions can be true without everything being true have become popular, largely under the influence of Priest's dialethism (e.g. 1987). I will sometimes allow myself the convenience of failing to distinguish implication of contradictions from logical triviality, but with the understanding that the claims thereby made could always be put in terms of triviality on grounds discussed in §8.4.

[4] In this chapter I set out the central aspects of my approach to the paradoxes. 'Inconsistency' views have been seen before, as noted. I haven't burdened this already long Chapter with extensive comparison of my views with the views of others, but I do undertake such comparison elsewhere.

8.1

Consider Grelling's paradox of heterological terms and Russell's paradox of the village barber. In the first we note that 'does not apply to itself' is a predicate that applies exactly to predicates that do not apply to themselves, while in the second we note that according to the decree of a certain village council, the village barber shaves all and only the men of the village who do not shave themselves. Contradictions appear to follow in both cases given that 'does not apply to itself' is itself a predicate and that the village barber lives in the village, for, given just this, 'does not apply to itself' applies to itself if and only if it does not and the barber shaves himself if and only if he does not; from there minimal commitments will get one an explicit contradiction.

Given the notable similarity of form—both 'paradoxes' involve the idea that a given object *x* bears a certain relation *R* to something *y* in a certain domain just in case *y* doesn't bear *R* to itself, and in both the problems arise when *x* is in the domain and we ask whether it does or does not bear *R* to itself—one might expect that both paradoxes would strike us as equally problematic, but they do not. We can see the crucial difference in perspicuous form by looking at the treatment of the Grelling as a diagonal argument in Simmons (1993). The one-place predicates of a language, English, for instance, could be listed, and hence one could set out a two-dimensional array with the same list of predicates on the top and the side; we write 't' in the relevant cell if the predicate at the side applies to the one on the top, 'f' if it does not. He continues:

> We obtain rows of t's and f's. {when it comes to the row for 'is monosyllabic'} each t in this row belongs to a column headed by a predicate in the extension of 'monosyllabic'. For instance, 'long' and 'new' are in the extension of monosyllabic. Consider now the diagonal sequence of t's and f's . . . Each t in this sequence corresponds to a predicate true of itself . . . and each f in this sequence corresponds to a predicate false of itself . . . Now form the antidiagonal sequence {that is, the sequence that swaps t's for f's in the diagonal} . . . This antidiagonal sequence cannot occur as a row {since it is different from every row at at least one place}. Now suppose there were a predicate of English that was true of exactly those predicates false of themselves. The row associated with this predicate would be exactly our antidiagonal sequence. So, by the diagonal argument, there is no such predicate. But in English there *is* such a predicate, namely 'heterological'. And we have a paradox.
>
> (17–18)

The paradox, then, is that we seem to be able to *prove* that English could not possibly contain a predicate that, as the penultimate sentence insists, it certainly appears to contain.[5] This is not our reaction to the barber; when we see that what has been

[5] It is an obfuscation to claim here that the problem in the argument is that the set of English predicates is not well defined, since the point is that *whatever* they are, we can give an argument that the

'decreed' is contradictory given the facts, we simply think 'so much the worse for the village council, decreeing does not make it so'.[6] Notice, however, that Simmons has no other argument for the claim that English contains a heterologicality predicate than appeal to the reader's credulity, and that he does not need one to seem convincing: speakers can be counted on to accept that the predicate is part of English. If this were not so we would simply have a theorem to the effect that English does not contain its own heterologicality predicate, rather than a paradox.

The trick, thus, in presenting the semantic paradoxes is to exploit the audience's impression that certain expressions exist. Here it is in Martin's introduction to Martin (1984):

> The Liar paradox has intrigued and frustrated philosophers since the fourth century BC. The problem is this: there are good reasons to accept as true the following two claims, *and* their incompatibility! The first claim is:
>
> (S) There is a sentence that says of itself only that it is not true.
>
> . . . It is hard to reject (S) in the face of a sentence such as 'This very sentence is not true' or 'What I'm saying now is not true.' There is no claim that such a sentence has a truth-value, or is non-deviant in any particular way—just that it has the reading it seems to have, or seems to be able to have. Even those who claim that such sentences are nonsense understand them well enough to see the need to make some such claim (1).

The second claim is simply that 'any sentence is true if and only if what it says is the case'. Since, then, what 'this very sentence is not true' says is that it itself is not true, it is true if and only if it is not. There would be no paradox here if the hearer did not accept that there are sentences that 'have or seem to be able to have' such readings. Here, too, however, the 'good reasons' for accepting (S) are nothing other than that it is 'hard to reject' it given that such sentences can be trotted out and are obviously grammatical English.

Taking the T-sentences, sentences of the form '*s* is true if and only if p' where what is substituted for 'p' is or translates *s*, all to be true therefore seems simply incompatible with the apparent presence of various problematic sentences in a language: such sentences appear to state what has to be the case for the sentences they mention to be true, yet contradictions can quickly be derived from them when the sentences

heterologicality predicate cannot be among them. It is thus part of my view to disagree with some of Simmons' comments on 'bad' diagonal arguments.

[6] And why not 'diagonalize' on the barber, concluding that the village barber does not live in the village, as a referee suggests? Because the mere assumption that there is a barber, and that the council has issued the decree as stated, does not imply that he does not live in the village: it implies only that either the barber does not live in the village, *or* that what the council decrees should be the case cannot be the case.

they concern include sentences of the sort Martin mentions. This raises significant questions for the familiar idea that the meaning of a sentence consists at least in part in truth conditions: such sentences appear to be meaningful, but what would appear to be accurate statements of their truth conditions imply contradictions given other uncontroversial claims. Thus, if we accept that the sentences of languages that contain expressions of the sort that Simmons and Martin discuss *are* meaningful, it would seem either that meaning does not consist in truth conditions after all, or at least that some sentences have rather unobvious truth conditions, truth conditions they themselves cannot be used to state.

8.2

Nevertheless, the idea that meaning consists at least in part in truth conditions has a good deal to be said in its favor. The prima-facie happiness of the schema MT:

If*s* means that p then *s* is true if and only if p

indicates as much.[7] Taking this impression seriously, the theory of meaning for English will be at least in part the attempt to generate, from a finite set of axioms, *all* of the T-sentences for English that the evidence indicates speakers of English accept. This has been the central point of the Davidsonian program all along, and is accepted by many who disagree with Davidson on various more specific points (see here the early essays in Davidson 1984).

That the finiteness of the set of axioms from which the T-sentences are derived is important is due to learnability considerations stressed by Davidson and others (e.g. Davidson 1984, 3–15). It is also important that the axiomatic structure matters: such theories must, among other things, generate good predictions about partial understanding of the following sort: if a speaker who does not understand 'cats meow', 'cats chase mice', 'cats like catnip', and so on comes to accept the relevant lexical axiom (viz. 'cats' refers to cats), she will come to understand those among these sentences for which she accepts the other relevant lexical and grammatical axioms. So merely generating the right consequences does not exhaust the task of the theory: discerning structure in how these consequences are generated is also a task of the theory so construed.

By far the majority view at least since Davidson's seminal papers and probably before is that what is stated in the axioms of such a theory is in some sense *known* to speakers

[7] See Soames (1999, 105) on the importance of MT. Elsewhere I discuss views such as that of Ludwig (2002) on which the paradoxes show that we should give it up.

or at least in principle *knowable*.[8] The assumption seems natural: if understanding a language is knowing what its sentences mean, then in stating a theory of meaning that is to generate the T-sentences for English, we are stating something in some sense known by speakers. Call this, following Pettit (2002), the *epistemic* conception of understanding. Pettit provides a nice survey of the epistemic conception as well as a good criticism, one based on examples that show that understanding, unlike knowledge, does not give rise to Gettier cases. I think Pettit's examples convince, but I am out to undo a view that even Pettit does not question. It is an assumption of the claim that understanding English is a kind of knowledge that what is known is at least *true*.[9] If this is so, in stating a semantic theory for English we are stating a theory that has to *be* true. This *factive* conception of understanding then sets a task for the theory of meaning: it must be possible, compatible with the evidence about what speakers think and do, to state in a true theory that which speakers grasp when they understand the language.

If this is right, however, we are in a great deal of trouble, for as we have seen the paradoxes appear to indicate that what speakers are inclined to accept about English cannot all be true. Speakers of English are inclined to accept that a subject–predicate sentence is true if and only if the object denoted by the subject has the property expressed by the predicate, that is, that ⌜Fa⌝ is true if and only if Fa; they are inclined to accept, for instance:

'snow is white' is true if and only if snow is white

Speakers of English likewise are inclined to accept that a predicate applies to all of the things that have the property it expresses, that is that for all *x*, ⌜*F*⌝ applies to *x* if and only if Fx; they are inclined to accept, for instance:

'red' applies to red things

The latter tendency, however, given 'does not apply to itself' on its natural reading, inclines speakers to accept the logically false:

'does not apply to itself' applies to itself iff it does not apply to itself

while the former, where Liar is 'Liar is not true', disposes them favorably toward

'Liar is not true' is true iff Liar is not true

[8] For just one example, the essays in the influential Evans and McDowell (1976) often clearly assume this; e.g. Foster (1), Davidson (34), Dummett (70). See also Pettit (2002) for a good survey with references.

[9] That Pettit does not question this is clear from his saying that (a) understanding is a state with propositional content and (b) that speakers can come to know 'facts' about the languages they understand. Both claims can be found at (2002, 549) See also Pettit (2005, 69) where the attitudes in question are explicitly stated to be true.

which in conjunction with this identity likewise implies a logical falsehood. This suggests that whatever our relation is to what is stated by the axioms of a theory of meaning for English, it cannot be knowledge or any other factive attitude such as true belief.

What is right in the idea that understanding is knowledge is that understanding is a cognitive state that explains various sorts of behavior and achievement. It is because speakers understand English that you can get them to believe things by saying things in English, get them to do things by giving them commands in English, and so on. The theory of meaning is thus part of an overall empirical study of what speakers do.[10] However, if we set out to state empirical theories that help us understand what people do and we find it helpful to conceive of some of the states that people are in as states that involve cognitive relations to theories, we need not in general assume that the theories are true. I will argue that, in fact, the theory according to which speakers understand English is not true, and that understanding English is thereby not *knowing* anything. I will argue, instead, that understanding English is a state that inclines speakers of English to believe falsehoods.

For the bulk of the chapter the state of ordinary speakers will be the main focus. However, since I am a speaker of English and I clearly do not believe the theory I discuss, this will raise questions both about how according to my own view I can express my own view in English and about how according to my own view I continue to understand English given that I disbelieve this theory. The resolution of these problems will come in the penultimate section: the full form of the view to be offered here will be that understanding English is being in a kind of sub-doxastic cognitive state such that, when one is in it, it seems to be the case that English sentences have certain semantic properties, that is, understanding a natural language is having a cognitive module that processes in accord with a logically false theory.[11] Ordinary speakers take these appearances at face value and hence believe what seems to them to be the case on particular occasions about the semantics of their natural language most of the time. I and readers who believe me do not accept these appearances, but we remain in this sub-doxastic state, much as those subject to familiar visual illusions fail to believe that things are as they seem while still undergoing the visual processes that give rise to the illusory visual appearances (see the discussion of Pettit 2005 below); hence we continue to understand English even though we do not believe the theory.

[10] The idea is hardly unfamiliar. See Davidson (1990) for one concise statement of the overall integration of the theory of meaning into a broader account of mind and behavior.

[11] This, of course, is an empirical claim, but that understanding is modular is well supported and familiar from work in the tradition of Chomsky (e.g. 1986). I will simply be adding the claim that the computed semantic theory is false.

8.3

The semantic paradoxes arise when arguments that would otherwise be accepted as demonstrating that English does not contain certain expressions run up against the impression that these expressions are part of English. At this point suspicion nearly always falls on the arguments: it is thought that since English obviously does contain the expressions, something must go wrong in the arguments. This ought to have struck us as odd, since the arguments are not complicated (Azzouni 2003, 342) and as applied without essential modification in other cases (e.g. in proofs of formal indefinability results or in the 'paradox' of the village barber) we simply accept that they show certain expressions or certain barbers not to exist. What I suggest is that we take the other route: let the arguments stand, accept that they show that competent speakers are inclined to accept something false, and try to understand what follows. Implicit in my way of going about things is a thesis about what sort of problems both the basic and revenge paradoxes pose: they are *empirical* problems about how to reconcile the evidence we have about the semantics of natural languages—evidence that comes from the impressions of what they are like that competent speakers share—with what various formal results, hailing from Tarski and Gödel, indicate cannot be the case.

Keeping in mind, then, that natural language semantics is a contribution to the empirical study of speakers and the role of linguistic communication in their lives, we can start by noticing that communication simply does not require that when someone says something and someone else understands it, the two both have *true* beliefs about the meaning of the sentence employed; it is sufficient, rather, that the two have *the same* beliefs about its meaning.[12] Suppose you and I both think that 'dog' means *cat*. Suppose also that we are otherwise ordinary, competent speakers of English. Suppose you believe that cats meow and that initially I do not. Given your beliefs about 'dog' and what seems to you to be the case about other expressions and constructions of English, you will be under the impression that 'dogs meow' is true if and only if cats meow. You therefore say 'dogs meow'. Since I, like you, believe that 'dog' means *cat* but am otherwise a typical speaker of English, I, like you, am under the impression that 'dogs meow' is true iff cats meow. Suppose also that I believe that what you say is true. Given what you have said, namely 'dogs meow', I will detach the right side of the biconditional and come to believe that cats meow. Hence you get me to believe what you believe, namely that cats meow, in a way that essentially but unproblematically involves beliefs about what the expressions we use to communicate mean that would be rejected by just about any speaker of English.

[12] Note that I do not say 'necessary'. See §8.7.

The basic point here is that since communication on the basis of accepted T-sentences involves the speaker moving from the right to the left of a T-sentence and the hearer moving back again, as long as the two accept the same T-sentences any putative mutual 'error' about them will be 'cancelled out'.[13] It is intuitive to us, given what we tend to accept as speakers of English, to call such 'erroneous' beliefs *false*. Note, however, that a theory that is intended to explain what speakers do with language needn't take any stand at all as to whether or not these attitudes are truly held: it need only say that they are shared.[14] The assumption that what speakers believe about the semantic properties of the expressions of the language they speak is true as opposed to false does no empirical work. Hence, an empirical account of the semantics of natural language is free to set this assumption aside.[15]

The point, then, is that understanding a language need not be *knowledge* of anything. As long as speakers operate with the same semantic theory communication goes fine.[16] Usually, of course, trying to come to know and trying to agree with others bear an important relationship; think here of how much of what we know we know by testimony. The connection should nevertheless not be overstated. With the semantics of natural language, I suggest, knowing and agreeing come apart completely: agreeing with others in the semantic theory in accord with which one processes linguistic input is one thing, while knowing this theory, even tacitly, is quite another. The idea that when people have shared attitudes about a social artifact such as a language these always 'constitute' a realm of 'facts' about this artifact which these people thereby

[13] If I know that you differ from me in some respects, I might take account of this. If I am standard and you are deviant as in the above example, I might hold that 'dogs meow' is true *as said by me* iff dogs meow, but that 'dogs meow' is true *as said by you* iff cats meow. I could still form the right beliefs based on this, and we could still communicate. So the point, more articulately, is that we need to agree about the truth conditions of sentences as uttered by certain people (and of course we will need to account for context to handle indexicality, etc.) These complications do not affect the point in the text and are therefore suppressed there.

[14] Heck's concluding remarks in Heck (2004) are a clear case in point of the need to give up the claim that we *know* semantic theories that prove to be inconsistent. Heck proposes accepting that the theory that correctly characterizes English semantics implies logical falsehoods, but his insistence that we 'tacitly know' this theory makes the view look mysterious. Giving up the epistemic and factive conceptions provides the way out.

[15] I would also maintain that the story is compatible with the transmission of knowledge as well as belief. The epistemology of testimony is, however, a difficult topic and the details would distract us from our purpose here, so allow me to give just one easy example. Suppose we are reliabilists about knowledge and augment the above example so that you believe reliably that cats meow. As long as you reliably only say things that you are reliable about, I'll come to be reliable as to the fact that cats meow, too. Given reliabilism, I now know that cats meow.

[16] A referee kindly suggests that there are helpful comparisons here to Kripke's Wittgenstein (Kripke 1982) and Davidson's claims to the effect that there are no such things as languages as usually conceived (Davidson 1986). Though I don't think my view is to be identified with either, the comparisons are instructive and I may take them up elsewhere.

know is a source of confusion here. If, however, our mutually having the impression that certain semantic claims about the expressions of English are true constitutes some 'facts' about what the expressions of English mean, then some of these facts are 'contradictory' in that they can only *be* facts if some contradictions are true; as we saw in the previous section, contradictions can easily be derived from statements that competent speakers of natural languages will accept, namely, that expressions of natural language can be named or described therein and that such languages contain semantic expressions such as 'is true' and 'applies to itself'. If one does not like contradictions—or, more generally, logical triviality—one had better let go of the idea that there are any such facts.[17] Hence, even if there is something to the idea that shared attitudes can constitute socially instituted facts, such putative constitution is still hostage to the demand of logical non-triviality.[18]

There is a useful comparison here with an example in Chihara (1979). Suppose the clubs in the area are in the habit of not allowing their own secretaries to be members. Suppose these excluded secretaries decide to form their own club. They maintain that the condition of membership is that one is entitled to join the club if and only if one is a secretary of a club of which one is not entitled to be a member. All might go fine until the club finds a need to hire its own secretary. Then its members have to ask: may the secretary join or not? Given what they believe about the conditions of membership, if he may, he may not, and if he may not, he may. The new situation makes clear that not all of what seems to be the case as to the conditions of membership can be true, yet that those are the conditions of membership will seem to the members to be true *by fiat*. The members of the club are in a pickle. The point, however, is that we, as empirical theorists of clubs, are not in a pickle—unless, that is, we assume that what the members of the club believe about the conditions of membership just has to be true, and there is no reason we should: nothing more could be demanded of an empirical account of this club than that it say what the members believe about the conditions of membership. Similarly, I suggest, nothing

[17] Gross (2006) provides a number of arguments related to the one given here for the claim that the truth-theories to which we take 'semantic competence' to be a relation need not be true, though not with application to the paradoxes. I discuss Gross's view and its relation to my own elsewhere.

[18] In a slogan: though meaning is determined by use, not every use that *appears* to determine a meaning *in fact* determines a meaning. At this point, when presenting this material, I very often encounter the view that this cannot be right, since if several people share the impression that certain sentences have certain truth conditions, that *makes it the case* that those sentences *do* have those truth conditions 'for them'—in their dialect, or what have you. I cannot stress enough that this is the view that has to go, since, if we take it seriously, it forces us either to accept that ' "Liar is not true" is true iff Liar is not true' is *true* of the language of people who are under the impression that semantic claims that imply it are true, or to embark on the 'tempting strategy' discussed in §8.5. Problematic views held by others infect our own theories if we hold that they are true, even when they are views about the semantics of the language those others speak.

more could be required of an empirical account of speakers of English than that it state the semantic theory that the evidence indicates governs competent speakers' processing of English sentences.

If this is right, the following option is open to us: simply accept that understanding English is best represented as a relation to a logically false semantic theory. It would follow from this theory, of course, that every sentence is both true and false. However, unless belief and the other attitudes speakers bear toward this theory were closed under logical consequence, it would not follow that one would find speakers *believing* that every sentence is both true and false, or even acting as though this were the case.[19] This, I will argue, is actually the situation. Speakers process in accord with an inconsistent theory, but since communication does not require their impressions about what their sentences mean to be true, and since they do not believe all the consequences of what they believe or otherwise have the impression that all of these consequences are true, they get along just fine.

Speaking against this will be the idea that the expressions of English *just have* to mean *something*—if not exactly what they appear to mean, then at least something close to it. I reject this claim, for reasons already stated: the assumption that expressions of English mean what they appear to mean, so that speakers' attitudes about them are true, does no empirical work. If the assumption did have such work to do, we would be forced to cast about for a view on which expressions mean something close to what they appear to mean, yet something far enough from it that no contradictions are implied by a statement of these meanings. We are not, however, forced to cast about for this any more than in the case of Chihara's club we are forced to posit 'real' conditions of membership in the club about which the members of the club are mostly right. In both cases people act in some respects as though theories that are in fact inconsistent are true and everything that needs to be explained is explained by the fact that they do so. Call this, if you wish, an *error theory* of understanding natural language.

8.4

It is generally thought that Tarski's mistake in thinking that truth could not be defined for languages of a certain 'rich' character was due to his simplistic adherence to classical logic. In 'The Semantic Conception of Truth', for instance, after presenting a simple version of the Liar, Tarski writes, 'If we now analyze the assumptions which

[19] Azzouni (2003, 348) stresses related reasons that speakers do not believe that everything follows from everything in English.

lead to the antinomy of the liar, we notice the following' (1944, 348) and he then gives the familiar list of three:

> (I) We have implicitly assumed that the language in which the antinomy is constructed contains, in addition to its expressions, also the names of these expressions, as well as semantic terms such as the term '*true*' referring to the sentences of the language; we have also assumed that all sentences which determine the adequate usage of this term can be asserted in the language. A language with these properties will be called '*semantically closed*'.
>
> (II) We have assumed that in this language the ordinary laws of logic hold.
>
> (III) We have assumed that we can formulate and assert in our language an empirical premise such as the statement (2) which has occurred in our argument.
>
> It turns out that assumption (III) is not essential, for it is possible to reconstruct the antinomy of the Liar without its help. But the assumptions (I) and (II) prove essential. Since every language which satisfies both of these assumptions is inconsistent, we must reject at least one of them.
>
> It would be superfluous to stress here the consequences of rejecting the assumption (II), that is, of changing our logic (supposing this were possible) even in its more elementary and fundamental parts. We thus consider only the possibility of rejecting the assumption (I). Accordingly, we decide *not to use any language which is semantically closed* in the sense given.
>
> (1944, 348–9)

((III) is 'inessential' because the Grelling doesn't require it.) Tarski here suggests that we cannot do semantics coherently at all unless there is something that we can say about the language of study that cannot be said in it; this is the familiar 'ascent to a richer metalanguage'. A common reaction to the passage has been to think that we can save semantic closure by instead parting ways with Tarski and modifying classical logic after all. Since it is often thought that English obviously is semantically closed in Tarski's sense, the idea has been that some move of this sort just *has* to work—those who want to study the semantics of natural languages can hardly 'decide' not to consider them, and Tarski himself obviously did not 'decide not to use' English.

In a more formal vein, both in the early 'The Concept of Truth in Formalized Languages' and in later works such as *Undecidable Theories* (Tarski, Mostowski and Robinson (1953)) Tarski operates within the confines of classical logic. The proof of the indefinability theorem as he presents it depends on separating the sentences of the object language into the true and the false, and then noting that the possibility of diagonalization means that there will be sentences that are in both sets *if* the language contains a predicate that has as its extension the set of its own truths (Tarski (1983, 250), Tarski, Mostowski and Robinson (1953, 46–8))[20] On the assumption that every set definable in the object language is definable in the metalanguage, a contradiction

[20] Actually, it is Gödel codes of the language's own truths, but I will leave this tacit.

is thereby provable in the metalanguage; the conclusion, then, by reductio, is that the object language does not contain an open sentence satisfied by all and only its true sentences, and hence that it is not the case that every set definable in the metalanguage is definable in the object language as well. In answer to Russell's question about diagonal arguments (Russell (1903, 366–8)—why some prove theorems while others generate paradoxes—there is a theorem and not a paradox here because we have no antecedent reason to believe that the languages under consideration *do* contain predicates that have as their extensions the set of their own truths. After all, they are languages for doing mathematics and it is simply a result that they are not sufficient for expressing their own semantics. But we do tend to think that English contains a predicate that has as its extension the set of its own truths, namely 'is true in English'. This makes Tarski's view look rather intolerable when it comes to English and the great attraction to subsequent theorists has been to think that if we could get more sophisticated about the sets into which we sort the sentences of English we could consistently state the true semantic theory for English without ascent to a 'stronger' metalangauge.

I cannot discuss every possible approach based on this thought in detail here—especially not the ones that have not been proposed yet. It is familiar from the literature that such approaches are plagued by the 'revenge problem' mentioned in the introduction: the attempt to consign certain sentences to non-classical semantic categories runs up against further sentences that it is simply empirically implausible to claim are not in English, but that subject the accounts themselves to a further paradox if we admit that they are in English (see Burge (1984, 87 ff.) and Priest (1987, 29) for classic statements of the problem). The only choice, then, is between empirical implausibility and inconsistency and at the end of the day all theorists settle on the former, insisting in one way or another that there are *some* semantic facts about English that can't adequately be expressed in English. Nobody acts happy about this, but it is almost universally taken to be the price of a consistent view.

Since so many detailed surveys of this issue are available, I will not take space here to attempt another one.[21] I will, instead, just provide the following general characterization of the situation. Semantic complications intended to avoid the indefinability theorem can take two forms: one can either deny that truth and falsehood are mutually exhaustive, or that they are mutually exclusive. Most approaches fall into the first category (the most famous being Kripke (1975)), while the dialethism of Priest (1987) falls into the second and of course requires that the consequence relation for the object language is such that not everything follows from a contradiction, as even proponents of such moves accept that absolute triviality

[21] A few good places to start: Priest (1987, ch. 1), Gupta and Belnap (1993, ch. 1), Simmons (1993, chs. 1–4), Soames (1999, chs. 5–6), Field (2002), (2003), and Scharp, this volume.

(every sentence is true) is unacceptable. The first strategy, however, gives rise to a strengthened liar ('this sentence is not true', 'this sentence is either false or meaningless', and their more complicated relatives for more complicated approaches)[22] unless the object language doesn't have an open sentence that is satisfied by all and only sentences that aren't true, while the second remains powerless against the Curry, which makes for absolute triviality without a detour through negation inconsistency, given just a sentence like:

(C)If C is true, then God exists.

plus a few applications of modus ponens and conditional proof (Whittle (2004) provides a concise treatment).

In any case we end up with extremely unhappy claims. We can only avoid open sentences that are satisfied precisely by sentences not in the set of truths by disallowing a definable 'exclusion' negation, whereas in English we take it to be rather easy, using expressions like 'not' and 'it is not the case that', to produce things we cannot say from things we can and vice versa, thereby treating English as containing an exclusion device. Likewise, there is an emerging consensus that we will never avoid the Curry if we do not give up the view that claims of the form '$(A$ and $(A \rightarrow C)) \rightarrow C$' and various close relatives are all true; approaches as different as Priest (1987) and Field (2002) and (2003) agree on this. But, of course, in whatever language we use when we say that schema has false instances, we treat 'if A and $(A \rightarrow C)$, then C' as just fine as long as we merely want to say that modus ponens is valid for '$\rightarrow$'. If English is supposed to be the object language of such views, so that conditionals in English have the semantics attributed to '$\rightarrow$', it turns out that there are truths about the semantics of English—namely, truths about what follows from what in English—that English itself is powerless to express: the fact that '$\rightarrow$' doesn't translate the metalanguage's 'if . . . then' is, after all, precisely what blocks the Curry.

Rather than directly taking on the army of philosophers and logicians committed to escaping the prima-facie import of the above claims, I simply want to offer something different: the claim that understanding natural language can be, and in fact is, a matter of processing sentences in accord with a logically false semantic theory. If we give up the factive conception of understanding we can simply accept that the standard indefinability results apply to English just as well as to any 'formalized' language. This, in turn, will allow us simply to accept the paradoxes' rather clear demonstration of the fact that the theory processing according to which is understanding English *cannot* be true. What appears to speakers to be the case when they understand English is

[22] I include contextualist approaches (e.g. Simmons (1993)) and their problems with appropriate variants 'this sentence is not true at any context' here. Michael Glanzberg's version of contextualism strikes me as far more promising than Simmons', but I must leave discussion of it for another occasion.

'ineffable' not in the mysterious sense that it is true but inexpressible in English, but in the much more straightforward sense that it is false.[23]

8.5

A tempting strategy at this point will be to say something like the following. Speakers of English are committed to many T-sentences. For instance, what seems to them to be the case about English implies:

(1) 'Snow is white' is true iff snow is white

and

(2) 'Liar is not true' is true iff Liar is not true.

The paradoxes show that something they are committed to is in fact false. Since (2) looks like one of the culprits, while (1) looks fine, this means that (1) is true and (2) is false. So, though the overall theory that speakers of English believe entails falsehoods, we can separate out a sub-theory that implies only truths: this will include (1) and exclude (2). Furthermore communication happens, really, in the fragment of the language governed by the consistent sub-theory.[24]

Appealing as it might be at this first sight, this strategy gets us into the business of distinguishing 'snow is white' from Liar: the former, according to it, has a status that

[23] The view I advocate could come in a stronger form than the one I actually maintain. One would have it in this form if one held, following Tarski himself (1983, 187), that the T-sentences for a language defined truth for that language, for on this combined position the view would amount to the claim that the concept of truth was itself inconsistent. (I leave it to proponents of such views to sort out what that might mean; see for instance Eklund (2002).) This might be a defensible position as well, and if I believed that the T-sentences defined truth, I would probably maintain it. I do not, however, believe this (see Patterson (2002)); I maintain, rather, that speakers grasp the concept of truth independently of the T-sentences for their language, and then have inconsistent beliefs about the conditions under which sentences of their language fall under it (Davidson (1990) shares my views on the relation between truth and the T-sentences). For the purposes of this Chapter, however, one can take me to be exploring an idea common to the view in both forms, the idea that speakers of English and other natural languages process in accord with an inconsistent theory.

[24] This thought comes in countless forms, many of which I have heard in conversation when presenting this material. It is generally the first thing that occurs to people who think about language but view the paradoxes as an annoying distraction. It is often expressed using comparisons to 'singularities', 'dividing by zero', or the idea that language use is a kind of know-how and we simply do not know what to do with such sentences. All such views commit their proponents to sorting the sentences of natural language into those that are not problematic in the way that paradoxical sentences are and those that are, and in addition to being subject to objections based on contingent paradox they introduce the standard revenge dialectic.

makes its T-sentence unproblematically true, while the latter has a status that makes its T-sentence untrue. But this latter 'third status' cannot be truth or falsehood, and so we are on the road to the revenge problem. What follows is that our overall theory of understanding and communication in English had better not get us into the business of distinguishing 'snow is white' from Liar—any attempt to do so will amount to some form of an orthodox response to the paradoxes, one which will inevitably generate either inconsistency or untruth in our own theory or 'ascent to a stronger metalanguage' and thereby empirical inadequacy when the object language is supposed to be English.

It is a very good thing, then, that in order to explain what we want to explain when we do empirical semantics for English we don't need to classify particular T-sentences or speakers' attitudes toward them as being true, false, or something else. Speakers are in a cognitive state of processing English sentences in accord with the axioms of a certain theory. We can explain what competent speakers of English *do* by appeal to their relation to these axioms. But none of the explaining we want to do depends on sorting these axioms or their consequences into the true and the false, since successful communication depends only on *shared* impressions of meaning, not *true* ones. It remains the case, also, that there is still something to choose empirically between theories with the same consequences: accurate predictions about partial understanding still need to be generated, so we need to get the compositional structure speakers attribute to the sentences of English *right* in order to account for what actual speakers will believe about various sentences under various conditions, e.g. that speakers who come to process as though 'cat' refers to cats will come process 'cat' sentences as having have truth conditions involving cats. This empirical application explains why the theory, though inconsistent, cannot simply be represented as some arbitrary contradiction.[25]

At this point objections press. I am saying that understanding English consists in processing in accord with a logically false theory. Since the whole point is to allow an approach the paradoxes entirely within classical logic, it follows that the theory in question is absolutely trivial in that it implies everything.[26] In particular, then, it implies that every sentence has every truth condition, that everything follows from everything in English, and that everything in English is both true and false. Surely, one might think, my view could not be empirically adequate for English, since we do

[25] Here see also the remarks in Azzouni (2003) on the difference between implication and derivation.

[26] A referee asks why I do not just 'dump *ex falso quodlibet*' and make life much easier on myself. I do not because, among other things, it will not help, as shown by the Curry. I also think it is valid, though I will not be arguing for that here. The reason I nevertheless take my life to be easy in this respect is that speakers, though they are committed to things that imply the validity of *ex falso*, usually in fact ignore it—not that they should, but they do.

not accept that everything follows from everything or that every sentence has every truth condition. As Azzouni notes:

> Although natural languages *are* inconsistent, ordinary speakers *do not believe* that they are. In fact, as far as I can tell, almost *no one* believes that ordinary languages are inconsistent. And this is hardly surprising since the avoidance of contradiction is, as I said, a norm of ordinary language; and *this* means that it is *rational* for ordinary speakers to believe that ordinary languages are not inconsistent—despite the presence of evidence (e.g. liar's paradoxes) to the contrary!
>
> (348)

Defending the view that competent speakers process in accord with an inconsistent semantic theory for their language requires making sense of this fact without in effect giving up on the view entirely. Gupta and Belnap press the point well:

> It may be said (as Chihara (1984) does say in connection with *ex falso quodlibet*) that, even though there is a proof of God's existence from the T-sentences, it does not follow that it is *reasonable* to infer God's existence from the biconditionals. Chihara (1984), 226, writes that 'in "real life" situations, one doesn't simply accept blindly the logical consequences of whatever one may initially have reason to believe.' The suggestion is that although the T-sentences represent the rules governing the word 'true', we need not accept all the logical consequences of these rules. We need accept only those consequences that are 'reasonable'. The problem with this suggestion is that by putting the entire burden of the theory on the notion of 'reasonable inference'—a notion of which no theoretical or even intuitive account is given—it obscures completely the contribution of the T-sentences to the meaning of 'true'
>
> (1993, 15)[27]

I agree with Gupta and Belnap that the appeal to a further theory of 'reasonable inference', one that trumps the logic one supposedly accepts, is a non-starter. Here is what I have to offer instead. Understanding English involves speakers in processing in accord with the axioms of a certain theory. As we know from the semantic paradoxes, these axioms imply everything. However, most of the time, speakers do not believe that they do. When someone says to them 'cats meow' they apply the relevant semantic axioms for lexicon and grammar, come to hold that 'cats meow' is true if and only if cats meow, come to believe (or suppose, consider, etc.) that cats meow, and reason accordingly. They focus on the T-sentence for 'cats meow' and do not get themselves bothered about what is implied by other things that seem to them to be the case (e.g. about Liar) for 'cats meow'. The only 'theory of reasonable inference' here is, just as it should be, logic itself—helped along by the point that people don't believe everything implied by all of their linguistic impressions about the content of a given sentence most of the time. What saves speakers from incoherence in the beliefs

[27] I noted above that I don't accept the implication of the last sentence that the T-sentences do contribute to the meaning of 'true', but that this will not matter for present purposes.

they form on the basis of particular acts of communication is that they do not stray far from immediate application of what they accept, on the basis of their linguistic processing, about the semantics of the expressions used.[28] This differs from the appeal to 'reasonable' inference precisely in that it is no part of my view to maintain that speakers aren't *committed* to straying from such immediate application: they are as long as they take at face value their impression that English has the expressive resources it appears to them to have.[29] But that they do not allows them to maintain the impression, most of the time, that their take on English semantics is consistent. Only when they encounter the paradoxes is this belief shaken as it should be.

Speakers can communicate perfectly well in a language about the semantics of which what seems to them to be the case is inconsistent because these impressions and the beliefs formed in accepting them aren't closed under logical consequence. One might think that nevertheless it would be good for them to switch to a language with an underlying consequence relation on which everything that seemed true could really be true, but there is simply no reason to do this. Nothing would be gained, and a good deal would be sacrificed in added complexity or lost expressive power;[30] think here of the high computational price that would be paid for a paraconsistent consequence relation, or tracking extra truth-values. Speakers, who have enough trouble with classical inference, would be even worse off trying to track these validities. The fan of consistency might here retreat to simple bans on reference to sentences, but these too have high prices. As it is common to note, the main everyday use of 'is true' is to commit the speaker to something that the speaker herself can't produce or inspect.[31] Demanding that speakers not attribute truth at all, or at least that they attribute it only to things that have been thoroughly inspected for lack of vicious referential loops and so forth, would greatly hamper speakers' ability to express themselves in natural language. Given that paradox can depend on contingent fact (Kripke (1975, 55),[32] only brutal cuts in expressive resources would guarantee consistency. In light of all this, it is no surprise that the cheapest solution is the one we find implemented: go with an inconsistent theory, let the contradictions

[28] Setting aside the fact that their non-linguistic beliefs are probably inconsistent, too. The view here is intended to explain how we could consistently form beliefs on the basis of communication in an inconsistent language.

[29] This is the fundamental point of disagreement with Eklund (2002), (2005), who, like Chihara, balks at drawing the conclusion that speakers ought to conclude that every English sentence is true. In contrast, I maintain that given what speakers accept about English semantics, they ought to accept that every English sentence is true; they are, however, saved from total incoherence by their logical lethargy. This difference gives rise to further disagreements with Eklund, all of which I discuss elsewhere.

[30] Note that expressive power here, like translation below, must on the view I present be understood in terms of the truth conditions that it will seem to competent speakers that sentences have.

[31] The point is a staple of the deflationism literature; see Azzouni (2001) for a good recent discussion.

[32] Reference is to the reprint in Martin (1984).

flow, and get on in life by ignoring them most of the time. Of course one could combine the view presented here with any non-trivial logic one pleased; my point is simply that the approach offered is sufficient to resolve the paradoxes even given classical logic.

8.6

The theory that best accords with the evidence about how speakers of English process English semantics is such that anything true according to it is also false, and anything false is also true. Since nothing is both true and false, this theory has to be false. It is therefore no part of my view to assert that anything is true in English, or that anything is false in English. Parsons grasps the point and sees it as the source of an objection:

> [an] assignment of truth-conditions to sentences of English would have to satisfy some conditions of coherence, since it is not just an account of the beliefs of English speakers but an account of the conditions under which what they say is *in fact* true. But then it is hard to see how such an account could make English inconsistent
>
> (1983, 244–5)

If natural language semantics is required to account for the conditions under which what speakers say is *in fact* true in English, then a view of the sort I offer is no good: we can't *merely* say that we're accounting for what seems to speakers to be the case about the conditions under which the sentences of English are true and leave it at that. Accepting Parsons' suggestion, though, unless the revenge problem is soluble, will inevitably make one's own account of English inconsistent, or will require saying that there are semantic facts about English that English is powerless to express, something that speakers of English fail to believe. I thus go with something close the suggestion Parsons here passes over: an empirically intended assignment of truth conditions to sentences of English *is* just an account of the theory in accord with which speakers process English sentences.

Nevertheless, one surely wants to account for the fact that in English we *try* to say true things, and that many things appear to be true or false *only*. If we are to take these impressions seriously, the view that 'snow is white' is simply true had better amount to something other than the claim that the facts about snow *plus* the semantic theory according to which speakers process in understanding English *determine* that it is true, because *they*, taken seriously, also determine that it is false.[33] Since nothing is both true

[33] Dialethists take note: I could reformulate in terms of the Curry and point out that the theory speakers believe implies that everything is true.

and false, we had better not take the theory seriously in that way. If, therefore, one has serious business that one takes to require saying things that are *actually* true as opposed to things that most speakers will merely *believe* to be true, one had better translate whatever one wishes to say into a consistent language, a 'rational reconstruction' of some relevant part of English. Translation on the present view is, of course, a mapping from expressions to expressions such that the semantic properties competent speakers take each to have are the same.

The view here is simply that if you want to express your belief that snow is white with a sentence that is *simply and unproblematically* true in its language, for real, the sentence in question had better not be evaluated by the semantic theory for English as individuated on any empirically plausible account. If, on the other hand, you are not so zealous, you can get by with saying something that you can be confident will seem to everyone to be true or false only. Human beings have not flocked to speaking consistent formalized languages for just this reason: we can get other speakers to believe that we believe things in English because we can use sentences of English that other speakers will *take* to be true only under certain conditions, because these other speakers will ignore most of the consequences of their own conception of English semantics when forming beliefs about the conditions under which a particular uttered sentence is true. Since speakers of English so reliably treat things said *as though* they had been said in a consistent language, practical purposes will pretty much never require retreat to a consistent *Begriffsschrift*.

If we want to think clearly about the paradoxes we must nevertheless keep the ideal of translatability into a consistent language in mind in order to be sure that the inconsistency of English is not being exploited. This applies especially to inconsistency approaches to the paradoxes themselves. The obvious question therefore concerns whether all of the claims of this chapter could be translated into a consistent language. Now the actual claims of this chapter most certainly could: I do not even need a truth predicate that applies to my own sentences in order to express my view. The more important question is whether in a consistent language I can fully state the claim that speakers process in accord with a certain theory which I then go on to state in full. Here is what looks to be the fly in the ointment. Will I be able to do so with the claim that speakers of English process truth conditions of English sentences in accord with a theory T, where what goes in for 'T' is a truth-conditional semantic theory that is empirically responsive to the evidence provided by speakers of English? Not quite. Suppose I try. Part of what I'll need to say is a translation in a consistent language of this:

Speakers of English process according to a theory on which 'does not apply to itself' applies to something if and only if that thing does not apply to itself.

If my putatively consistent language can itself be understood, it itself is going to have a finitely stateable semantic theory. Unless I want to ruin everything by insisting that my attitude ascription performs a magic trick, this semantic theory is going to need to generate the truth condition of the ascription from an axiom that tells me the truth conditions for instances of 's processes according to a theory on which p' based on things that the semantic theory says about what's substituted in for 'p'. In this case, that is:

> 'does not apply to itself' applies to something if and only if that thing does not apply to itself.

The problem here isn't that I am *asserting* this in my consistent language: I am not. But my language is going to have to have the resources to *express* it. So it will have to have a biconditional and the resources to express the sentences on either side. But these include 'does not apply to itself'. So I'm going to need to handle 'does not apply to itself' in my metatheory for the regimenting language, and, more basically, 'applies', from which it is generated. But, then, if my metatheory is consistent, it will have to insist that its own 'applies' isn't in the range of application of the consistent regimenting language's 'applies'. So I'll have to admit that 'applies' in my regimenting language doesn't really have the application relation as its extension, and hence doesn't fully translate the English 'applies', since speakers of English are under the impression that 'applies' has the application relation as its extension. But then whatever I am saying in my consistent language when I say that speakers of English process according to a theory which I go on to *state*, I am not adequately expressing the state speakers are in. If I did adequately express it, on the other hand, a falsehood would be true—since if the language I used had the semantic features speakers of English take it English has, a contradiction would be true—and falsehoods aren't true.

A comparison, however, should make clear why there is nothing strange about this. Nearly everyone admits that first-order translations of English are imperfect. Nearly everyone understands, though, that various goals one might have in theory construction might be well served by stating things in a first-order language (e.g. one might want completeness). There is nothing wrong with accepting, in the service of other goals in theory construction, the unavoidable imperfection in translation involved in translating one language into another where competent speakers of the second take it to lack expressions that competent speakers of the first take the first to have. The situation is the same here. Ascribing the attitudes in which understanding English consists in a language that can itself be given a consistent semantic theory requires ascribing them in a language that lacks semantic features that speakers of English are under the impression English has, since its lacking them is a necessary

condition of the language's having a consistent semantics at all.[34] Hence it requires, unavoidably, imperfect reporting of attitudes that concern the semantics of English.[35] But this is no more unexpected than the loss familiar from first-order translations. It will also not impede the explanatory work to which the theory is put: saying that the speakers have the impression that 'is a dog' applies to dogs is going to get us very nice explanations of what they do with 'dog' sentences (in spite of the fact that *our* 'applies to' is not a perfect translation of their 'applies to', since we don't take it to have the whole of the extension that they take theirs to have) while saying that they have the impression that 'does not apply to itself' applies to all and only predicates that do not apply to themselves is going to provide an equally good explanation of why they are puzzled when they encounter the Grelling (in spite of the same facts), since it will explain why they cannot avoid the conclusion that the predicate applies to itself if and only if it does not, while they are also under the impression that this cannot be the case.

In work on the paradoxes it has largely been accepted, if grudgingly, that, one way or another, coherently stating a full semantic theory for a natural language like English requires use of a language into which English can be translated, but which cannot fully be translated into English. This allows the claim that English semantics is factual, but at the price of implausible claims about what is and is not in English. The strategy I propose is the opposite: recognize that natural language semantics is not factual, but is rather about the (as a body, false) attitudes speakers have toward natural language, and accept that the theory is to be stated in a language that can be translated into English, but into which English cannot be fully translated.[36]

8.7

A number of related questions as to how I can state my own view will press at this point. First, if things are as I say they are and what seems to speakers to be the case about the semantics of natural languages is false, how can I expect to get my point across in English? On a related point, if I do not believe the theory I say speakers of English believe, how can I understand anything anyone says to me in English? How,

[34] I don't suggest here that there is a unique combination of logical and expressive resources to be used for the purpose; maybe a reasonable pluralism or pragmatism can govern regimentation, or maybe some further argument, not given here, could suggest a best candidate.

[35] Assuming, as I have been, that using an imperfect translation of the sentence the subject would use to express an attitude in an ascription of that attitude results in an at least somewhat inaccurate report of the attitude. Perhaps this isn't so, but then my case is only strengthened.

[36] Again, keep squarely in mind that translation is preservation of what semantic properties competent speakers will *take* expressions to have, not preservation of semantic properties actually had by expressions.

that is, can I communicate in English if I don't believe the theory in accord with which competent speakers understand it? Second, and closer to home, how can I understand English at all (as, clearly, I do) if I don't believe this theory? In what does understanding consist, if not belief in the semantic theory shared with other speakers?

To get started, consider this idealized but basically accurate characterization of how ordinary speakers extract information from the utterances of speakers they take to speak truly on the basis of accepted T-sentences. Suppose that A hears B say 'snow is white', and suppose that A believes that what B says is true. A can reason as follows (for simplicity here as elsewhere I suppress distracting relativizations to English, places, times, etc.):

1. B says 'snow is white'.
2. 'snow is white' is true iff snow is white.
3. What B says is true.
4. So, 'snow is white' is true(by 1 and 3).
5. So, snow is white(by 2 and 4).

Now I, as part of my view, have lost the crucial premise (2): I believe that speakers of English process according to a theory that implies (2), but I also believe that this theory is false. So how can I extract the right information from anything anyone says (as, clearly, I still can)? Easily: I just, in the role of A, reason as follows:[37]

1. B says 'snow is white'
2. B believes that 'snow is white' is true if and only if snow is white.
3. So, B wants to say something true.
4. So, B believes that snow is white (1 through 3, assuming B is minimally rational).
5. What B believes (about non-semantic topics) is true.
6. So, snow is white (by 4 and 5).

I can perfectly well extract information from the utterances of people I trust on other topics even if I don't accept their beliefs about semantics at all: I need simply take note of the fact that *they* believe their utterances to have various truth conditions and that thereby their utterances express their beliefs that these truth conditions obtain. This is exactly what I do: I don't believe that what typical speakers of English accept about English semantics is true, but, knowing that they accept it, I can still discover what they believe about non-semantic matters based on what they say and thence draw conclusions about how things are. I can likewise convey belief to hearers who I know will take my utterances to have certain truth conditions by applying what seems to them to be the case about English semantics: knowing that someone is in a state that disposes her to believe that 'snow is white' is true iff snow is white, I can, assuming

[37] I thank Thomas Hofweber for encouraging me to clarify the reasoning here.

she trusts me, convey the information that snow is white by saying 'snow is white'. I exploit beliefs I do not share in order to communicate with speakers who accept a semantic theory I deem false. Accepting the present view is thus no impediment to communication.

So what sort of state is understanding English? I have said that competent speakers of English tend to believe falsehoods about English semantics, and that their ability to communicate in English is based on their impressions as to what things mean being shared rather than their being true. This might motivate the claim that understanding English is a matter of falsely believing a certain semantic theory for English. This, however, would have the rather odd consequence that I who do not believe this theory do not understand English. A better route is to adopt a suggestion from Pettit (2002, 2005) that goes back to Chomsky (e.g. Chomsky (1986)): take understanding itself to be a *modular* or *sub-doxastic* cognitive state that gives rise to (quasi-) perceptual impressions that certain things are the case. Understanding English is thus being in a state such that it seems to one as though various things are the case, e.g. that 'snow is white' is true iff snow is white, and that 'Liar is not true' is true iff Liar is not true. As with other such capacities, one has a prima-facie inclination to believe that things are as they seem them to be, an inclination which can, however, be overridden.

Typical cases of visual illusion such as the Müller–Lyer provide a good comparison here.[38] In virtue of the capacity for sight, it seems to one as though various things are the case. In the ordinary run of cases one, on this basis, believes these things to be the case. However, in the illusion two lines that are actually of the same length seem visually to differ in length. One who knows the trick will thereby not believe that what visually seems to be the case really is the case, yet, for all this, it will continue to seem visually to be the case. Hence the capacity to have it seem to one visually that various things are the case is not sufficient for, and does not consist in, belief that these things are the case.

It is similar, I claim, with understanding a natural language. The combination of innate dispositions and the promptings of experience put one into a sub-doxastic state such that various things seem to be the case—that, for instance, 'snow is white' is true iff snow is white and 'Liar is not true' is true iff Liar is not true. In the ordinary run of cases, one believes the immediate deliverances of this semantic capacity and reasons accordingly: this is what happens among ordinary speakers most of the time. However, thorough reflection on the paradoxes indicates that what seems to be the case in understanding simply cannot be the case. Yet, I claim, one remains in the

[38] Pettit (2005) makes use of the example, though not in the service of the view that understanding can be a relation to a false theory, but merely in service of the view that one might understand a language without believing that it has the semantics it seems to have.

relevant sub-doxastic state. One is thus in the state of sub-doxastically 'cognizing', to use Chomsky's term (1986) a theory that one nevertheless believes to be false.[39]

This, then, is my full view: understanding a natural language is sub-doxastically cognizing a semantic theory that the paradoxes show to be logically false. Ordinary speakers believe the impressions generated by their processing in accord with this theory but get along just fine because they don't believe all the consequences of what also seems to them to be the case about English semantics. I and those who accept my view do not believe this theory, but we can still communicate with those who do because we can take account of the fact that *they* believe the appearances to which cognition of it subjects them, and we still understand our mother tongues because we are still in the sub-doxastic state in question.

8.8

Natural languages appear to defy the clear purport of the standard indefinability results: they appear to be richly stocked with semantic vocabulary and yet they also appear to be logically non-trivial in the sense that it is possible for some but not all of their sentences to be true. If we take this impression at face value, our task in finding a solution to the paradoxes is to explain how a language could possibly have both of these features; we need to discover the underlying logic that allows this amazing feat to be pulled off in defiance of settled logical results. The ongoing research program, fuelled largely by the dynamic of semantic 'revenge', is to discover a logic that fits the bill.

What goes unchallenged in all of this is the assumption that natural languages do manage to be rich in semantic vocabulary yet logically non-trivial. The answer, and the way out of the real problem, is to recognize that this is the assumption that misleads us. Speakers of natural languages process according to theories on which they are expressively rich and, since they tend to accept these impressions as true but not to accept all of their consequences, they believe these languages are expressively rich but logically non-trivial. However, there is simply no reason speakers have to be right about this: speakers—including philosophers of language—have misleading impressions about the expressive power a language can have because they process their own languages according to logically false theories.

The main reason that inconsistency views of the semantic paradoxes have been so widely rejected is that people have read them as claiming that the *true* semantic theory

[39] The view thus corrects Chomsky's by jettisoning commitment to the factive and epistemic conceptions of the state of the 'language module'. Works like Chomsky (1986) are shot through with unnecessary commitment to the epistemic conception.

of English is inconsistent, that is, that the true semantic theory of English is also *false*. It may well be that inconsistency theorists have themselves been confused on this point; in any case, in order to see our way out of views subject to this objection we need to realize that if semantics is construed as part of an empirical study there is no need for it to do anything more than state the theories according to which speakers understand the languages they speak. The inconsistency view thus needs to be formulated not as the claim that natural languages are inconsistent, but as the claim merely that competent speakers process them in accord with an inconsistent theory. Keeping clear that having impressions about what is the case that are shared by others is not necessarily knowing is, in turn, all that is required to see the thought that understanding a language can be a relation to a false theory. The payoff is significant: we have a coherent story about the paradoxes that sets aside the otherwise inevitable 'ascent to a stronger metalanguage' because we need not say that there is something true of the semantics of English that just 'comes out wrong' when we try to say it in English.[40]

References

Azzouni, Jody (2001). 'Truth via anaphorically unrestricted quantifiers'. *Journal of Philosophical Logic* 30: 329–54

——— (2003). 'The strengthened liar, the expressive strength of natural languages, and regimentation'. *The Philosophical Forum* 34: 329–50

Burge, Tyler (1984). 'Semantical paradox'. In Martin (ed.), *Recent Essays on Truth and the Liar Paradox*. New York: Oxford University Press: 83–117

Chihara, Charles (1979). 'The semantic paradoxes: a diagnostic investigation'. *Philosophical Review* 88: 590–618

——— (1984). 'The semantic paradoxes: some second thoughts'. *Philosophical Studies* 45: 223–9

Chomsky, Noam (1986). *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger Publishers

[40] Many people have given me helpful feedback on the material presented here. I thank audiences at the Université du Québec à Montréal, University of Ottawa, Queen's University, Ontario, The Society for Exact Philosophy meetings in Toronto, May 2005, the Canadian Philosophical Association meetings in London, Ontario, in June 2005, The Understanding and Communication conference in Lisbon, Portugal, June 2005, Logica 2005 in the Czech Republic, King's College, London, the University of Glasgow, and Tufts University. For helpful comments during those sessions I thank in particular Bradley Armour-Garb, Jody Azzoouni, David Bakhurst, Michael Detlefsen, Luc Forcher, John Hawthorne, David Hyder, Gary Kemp, Phil Kremer, Ernie Lepore, Guy Longworth, Peter Ludlow, Mathieu Marion, Adèle Mercier, Martin Montminy, Claude Panaccio, Graham Priest, Greg Ray, Mark Richard, Adam Rieger, Paul Rusnock, Barry Smith, Mark Textor, James Woodbridge, and many others. For comments on previous drafts and discussion in correspondence I thank Jody Azzouni, JC Beall, Matti Eklund, Michael Glanzberg, Bruce Glymour, and Kevin Scharp.

Davidson, Donald (1984). *Inquiries into Truth and Interpretation*. New York: Oxford University Press

_____ (1986). 'A nice derangement of epitaphs'. In Lepore (ed.), *Truth and Interpretation*. Oxford: Blackwell: 433–46.

_____ (1990). 'The structure and content of truth'. *Journal of Philosophy* 87: 279–328

Eklund, Matti (2002). 'Inconsistent languages'. *Philosophy and Phenomenological Research* 64: 251–75

_____ (2005). 'What vagueness consists in'. *Philosophical Studies* 125: 27–60

Evans, Gareth, and McDowell, John, (eds.) (1976). *Truth and Meaning*. New York: Oxford University Press

Field, Hartry (2002). 'Saving the truth schema from paradox'. *Journal of Philosophical Logic* 31: 1–27

_____ (2003). 'A revenge-immune solution to the semantic paradoxes'. *Journal of Philosophical Logic* 32: 139–77

Gross, Steven (2005). 'Linguistic understanding and belief'. *Mind* 114: 61–6

_____ (2006). 'Can empirical theories of semantic competence really help limn the structure of reality?' *Nous* 40: 43–81

Gupta, Anil, and Belnap, Nuel (1993). *The Revision Theory of Truth*. Cambridge, Mass.: MIT Press

Heck, Richard G. Jr. (2004). 'Truth and disquotation'. *Synthese* 143: 317–52

Kripke, Saul (1975). 'Outline of a theory of truth'. *The Journal of Philosophy* 72: 690–716. Reprinted in Martin (ed.), *Recent Essays on Truth and the Liar Paradox*. New York: Oxford University Press: 53–81. References are to the reprint

_____ (1982). *Wittgenstein on Rules and Private Language*. Cambridge, Mass.: Harvard University Press

Ludwig, Kirk (2002). 'What is the role of a truth theory in a meaning theory?' In Campbell, O'Rourke, and Shier (eds.), *Meaning and Truth: Investigations in Philosophical Semantics*. New York: Seven Bridges Press

Martin, Robert (1984). 'Introduction'. In Martin (ed.), *Recent Essays on Truth and the Liar Paradox*. New York: Oxford University Press: 1–8

Maudlin, Tim (2004). *Truth and Paradox: Solving the Riddles*. New York: Oxford University Press

McGee, Vann (1991). *Truth, Vagueness and Paradox: An Essay on the Logic of Truth*. Indianapolis: Hackett

Moore, G. E. (1942). 'A reply to my critics'. In Schlipp (ed.), *The Philosophy of G. E. Moore*. New York: Tudor Publishing

Parsons, Charles (1983). *Mathematics in Philosophy: Selected Essays*. Ithaca: Cornell University Press

Pettit, Dean (2002). 'Why knowledge is unnecessary for understanding language'. *Mind* 111: 519–49

_____ (2005). 'Belief and understanding: A Reply to Gross'. *Mind* 114: 67–74

Priest, Graham (1987). *In Contradiction: A Study of the Transconsistent*. Dordrecht: Martinus Nijhoff

Russell, Bertrand (1903). *The Principles of Mathematics*. Cambridge: Cambridge University Press

Simmons, Keith (1993). *Universality and the Liar: An Essay on Truth and the Diagonal Argument*. New York: Cambridge University Press

Soames, Scott (1999). *Understanding Truth*. New York: Oxford University Press

Tarski, Alfred (1944). 'The semantic conception of truth and the foundations of semantics'. *Philosophy and Phenomenological Research* 4: 341–75

Tarski, Alfred (1983). 'The concept of truth in formalized languages'. In *Logic, Semantics, Metamathematics*, 2nd edn., Woodger (trans.), Corcoran (ed.), Indianapolis: Hackett, 1983: 152–278

——— (1986). 'Der Wahrheitsbegriff in den Sprachen der deduktiven Disziplinen'. In *Alfred Tarski, Collected Papers Volume 1*. Boston: Birkhäuser

Tarski, Alfred, Mostowski, Andrzej, and Robinson, Raphael (1953). *Undecidable Theories*. Amsterdam: North Holland

Whittle, Bruno (2004). 'Dialethism, logical consequence and hierarchy'. *Analysis* 64: 318–26

Woodbridge, James (2006). 'Truth as a pretense'. In Kalderon (ed.), *Fictionalism in Metaphysics*. New York: Oxford University Press

9

Revenge, Field, and ZF

Graham Priest

9.1 Introduction

This chapter deals with three interrelated issues:

1. What is the 'revenge' phenomenon?
2. How does it bear upon Field's account of the semantic paradoxes?
3. Is the notion applicable to the set theoretic paradoxes?

The meanings of these questions, and the connections between them, will become clear in due course.

9.2 Revenge

Let us start with a statement of the simple liar paradox. This concerns a statment, L, of the form $F\langle L\rangle$, where F is the falsity predicate, and angle brackets indicate some name-forming device. If T is the truth predicate, then the T-scheme assures us that for any closed sentence, A:

$$T\langle A\rangle \leftrightarrow A$$

Substituting L, we have:

$$T\langle L\rangle \leftrightarrow F\langle L\rangle$$

Then applying the Principle of Bivalence for L, $T\langle L\rangle \vee F\langle L\rangle$, we infer $T\langle L\rangle \wedge F\langle L\rangle$, which contradicts the Principle of Univalence applied to L, $\neg(T\langle L\rangle \wedge F\langle L\rangle)$.

When people attempt to give an account of the paradoxes of semantic self-reference, such as the liar, they invoke certain machinery (a theory of truth, truth-value gaps, revision, etc.). It would seem to be that in many, if not all, cases this machinery can be deployed to formulate a related version of the paradox, just as virulent as the original. This is what I shall understand, for the purpose of this chapter at least, as the 'revenge' phenomenon.

There is, in fact, a uniform method for constructing the revenge paradox—or extended paradox, as it is called sometimes. All semantic accounts have a bunch of Good Guys (the true, the stably true, the ultimately true, or whatever). These are the ones that we target when we assert. Then there's the Rest. The extended liar is a sentence, produced by some diagonalizing construction, which says of itself just that it's in the Rest. The diagonal construction, because of its ability to tear through any consistent boundary, may then play havoc. This shows, incidentally, that the extended paradox is not really a different paradox. The pristine liar is the result of the construction when the theoretical framework is the standard one (all sentences are true or false, not both, and not neither). 'Extended paradoxes' are simply the results of applying the construction in different theoretical frameworks.

To see what options there are for handling the revenge situation, it is useful to look at it from the following perspective. The semantic paradoxes arise, in the first instance, as arguments couched in natural language. One who would solve the paradoxes must show that the semantic paradoxes do not, despite appearances, lead to contradiction (or, at least, triviality in the case of a dialetheic approach, but let us focus on consistent approaches for the moment; I will comment on the dialetheic case later). And it is necessary to show this for every concept in the semantic family, for they are all deeply implicated in paradox. Attempts to do this, given the resources of modern logic, all show how, given a language, $\mathcal{L}$, to construct a theory, $\mathcal{T}_{\mathcal{L}}$, for the semantic notions of $\mathcal{L}$, according to which they behave consistently. We now have a series of options.

Horn 1 Are the concepts of $\mathcal{T}_{\mathcal{L}}$ expressible in $\mathcal{L}$? If the answer to this is 'yes', it always seems possible to use the resources of $\mathcal{T}_{\mathcal{L}}$ to formulate the extended paradox, and so obtain a contradiction. Neither is this an accident. For since the concepts of $\mathcal{T}_{\mathcal{L}}$ are expressible in $\mathcal{T}_{\mathcal{L}}$, and since, according to $\mathcal{T}_{\mathcal{L}}$, things are consistent, we should be able to prove the consistency of $\mathcal{T}_{\mathcal{L}}$ in $\mathcal{T}_{\mathcal{L}}$. And provided that $\mathcal{T}_{\mathcal{L}}$ is strong enough in other ways—for example, if it contains the resources of arithmetic—then we know that $\mathcal{T}_{\mathcal{L}}$ is liable to be inconsistent, by Gödel's second incompleteness theorem. The upshot of this case is, then, inconsistency.

Horn 2 If the answer to our original question is 'no', the concepts of $\mathcal{T}_{\mathcal{L}}$ are not expressible in $\mathcal{L}$. In this case, we ask another question: are they expressible in some

other language? If the answer is yes, then $\mathcal{L}$ is expressively incomplete. There are certain semantic concepts that it cannot express. But then, in that case, the original problem of showing that our semantic concepts behave themselves has not been solved. For $\mathcal{T}_{\mathcal{L}}$ deals only with the semantic concepts of $\mathcal{L}$, and now there are others, also prone to generate inconsistency—as horn 1 of the situation shows—that have not been dealt with.

Horn 3 The other possible answer to this question is 'no': the concepts in question are not expressible at all. If this is to be a robust theoretical position, and not to lapse into gesturing at the ineffible, we must insist that the concepts in question are not meaningful; they do not exist. This is the case of inexistence (non-existence). At first blush, this position would seem to be shamelessly self-refuting, since the theorist has depended upon those very notions in giving their own account. But things are not quite so straightforward. We have talked simply of the concepts of $\mathcal{T}_{\mathcal{L}}$. But there are two ways in which a theory can invoke concepts. It may do so explicitly, by giving them names, reasoning about them, etc. If the semantic concepts of $\mathcal{L}$ are invoked in this way, then we do indeed have immediate self-refutation. But, more subtly, the concepts may not be invoked explicitly: they may be presupposed in some way, thus being invoked implicitly. If this is the case, denying the meaningfulness of the concepts is an option, if one can sustain the view that the concepts are not *really* presupposed. How problematic this move is now turns on how robust the presupposition is. If it is one that is hard to gainsay, the theory would appear to be in just as much trouble.

Just to illustrate this last possibility, think of Wittgenstein's *Tractatus*. The statements of the *Tractatus* say that a certain kind of sentence is meaningless (*unsinnig*). According to the *Tractatus*, most of the statements of the *Tractatus* are of this kind. But the *Tractatus*, though it does not *say* this, would seem to presuppose that these sentences *are* meaningful. It uses them. In the end, though, Wittgenstein simply denies the presupposition, and insists that they really are meaningless. I do not think that the move can be sustained coherently. But we need not go into that now. The situation illustrates how an account may presuppose something that it does not explicitly assert, how that presupposition may be denied, and the problems that this sort of move is wont to give.[1]

The upshot of the preceding discussion is this. The revenge scenario poses the theorist with three possibilities: inconsistency, incompleteness, and inexistence, each with its own characteristic problems. One or other of these horns must be selected and coped with.

[1] The Tractarian situation is certainly one of self-reference, though it is not a standard paradox of self-reference. Its structure is, however, one of an inclosure; so it is the same form of the paradoxes of semantic self-reference. For further discussion, see Priest (1995), ch. 12.

9.3 An Illustration

This is all rather abstract. Let me illustrate it with an example.[2] Consider the Tarskian solution to the semantic paradoxes. We start with a language, $\mathcal{L}_0$, with no semantic concepts. To reason about the semantics of $\mathcal{L}_0$, we move to a different language, $\mathcal{L}_1$, which contains the truth predicate, T_0, applying to the sentences of $\mathcal{L}_0$. That is, we have all instances of the T -schema, $T_0 \langle A \rangle \leftrightarrow A$, where A is a sentence of $\mathcal{L}_0$. To reason about the semantics of $\mathcal{L}_1$, we have to repeat the move, generating a hierarchy of language, which may be depicted as follows:

Language	T-**Schema**	**Legitimate Instances**
$\vdots$	$\vdots$	$\vdots$
$\mathcal{L}_{i+1}$	$T_i \langle A \rangle \leftrightarrow A$	$A \in \mathcal{L}_i$
$\vdots$	$\vdots$	$\vdots$
$\mathcal{L}_2$	$T_1 \langle A \rangle \leftrightarrow A$	$A \in \mathcal{L}_1$
$\mathcal{L}_1$	$T_0 \langle A \rangle \leftrightarrow A$	$A \in \mathcal{L}_0$
$\mathcal{L}_0$	None	

Given techniques of self-reference, it is easy enough to construct a sentence, L, such that $L = \neg T_i \langle L \rangle$. If we had $T_i \langle L \rangle \leftrightarrow \neg T_i \langle L \rangle$, and given that we have the Law of Excluded Middle, $A \vee \neg A$, we would have a contradiction. But we do not. L is a sentence of $\mathcal{L}_{i+1}$; so we have only $T_{i+1} \langle L \rangle \leftrightarrow \neg T_i \langle L \rangle$, and the liar-reasoning is blocked. More generally, given an appropriate formal definition of the hierarchy, and an interpretation for $\mathcal{L}_0$, one can prove that the hierarchy of truth theories is consistent.

Call the level in the hierarchy at which a sentence appears (or the first such level if the levels are cumulative, as they are usually taken to be) its *rank*. Let $rk(x)$ be the rank of x. The Good Guys in this construction are the ones that are true at their rank; that is, the sentences that satisfy the predicate $T_{rk(x)}x$. So the extended liar is one that says of itself that it is not true at its rank:

$$L : \neg T_{rk(\langle L \rangle)} \langle L \rangle$$

We now have the possibilities corresponding to the three horns.

The first is that the notion of being true at its rank is expressible in some language in the hierarchy. Suppose this is $\mathcal{L}_i$. Then L is a sentence of $\mathcal{L}_i$, and $rk(\langle L \rangle) = i$. So at level $i + 1$, we have $T_i \langle L \rangle \leftrightarrow \neg T_{rk(\langle L \rangle)} \langle L \rangle \leftrightarrow \neg T_i \langle L \rangle$, and we have contradiction. This is the inconsistency case.

[2] For further examples, see Priest (1987), ch. 2, and the 2nd edn., 19.3.

The second is that the predicate $T_{rk(x)}x$, though meaningful, cannot be expressed in the hierarchy. What this shows is that there are semantic concepts with the potential to generate contradiction, and which are not dealt with in the theory. This is the incompleteness case.

The third is the inexistence case. We can deny that there is such a concept as 'truth at its rank', that it is a meaningful notion. In the Tarskian construction, the concepts employed in its expression are invoked explicitly by the theorist; hence a denial of its existence (meaningfulness) is a simple self-refutation. (To specify the hierarchy, the theorist must say that for each level of the hierarchy, y, there is a truth predicate, T_y, at level y. So quantification into the subscript place of the truth predicates must be legitimate.)

Before we move on to Field, let me conclude with a word on dialetheism and revenge. In a dialetheic treatment of the semantic paradoxes, the Good Guys are the truths. The Rest are the things that are false but not also true (assuming for the sake of argument that there are no truth-value gaps). So the extended liar is a sentence, L, of the form $F\langle L\rangle \wedge \neg T\langle L\rangle$. Assuming all these concepts to be expressible in the language, reasoning about this sentence in the natural way one can demonstrate that $F\langle L\rangle \wedge T\langle L\rangle \wedge \neg T\langle L\rangle$.[3] This is a contradiction. So the revenge phenomenon applies just as much to the dialetheic theory. We end up in Horn 1, inconsistency. But clearly, though this horn of the dilemma is devastating for consistent accounts of the paradoxes, it is not so for dialetheic ones.

9.4 Field

With this background, let us now move to Field's account of the semantic paradoxes, and see what happens there. I will not give an exegesis of his account. It can be found in Field (2003), (2005), and (2007).

Since the question of what can and what cannot be expressed in a language is clearly crucial, let us start by getting clear what the language of Field, the theorist, is. It is essentially the language of *ZF* augmented by a truth predicate, T, and a non-truth-functional conditional, $\rightarrow$, to be deployed in stating the T-schema. Field gives a semantics for this language. According to the interpretations he defines, the purely set-theoretic vocabulary behaves classically; we may therefore reason as in classical *ZF* as long as T and $\rightarrow$ are not being used (as opposed to mentioned) in the reasoning. Generally, though, the semantics are many-valued, with the unique designated value, 1, which is the value of the Good Guys (the things that we can

[3] See Priest (1987), 2nd edn., 20.3.

correctly assert). The semantics can all be described in *ZF*, however, so it is permissible to reason classically about it.

In this context, the extended paradox is clearly generated by a sentence, *L*, of the form $Val(\langle L \rangle) \neq 1$, where $Val(x)$ is 'the value of *x*', and having value 1 marks out the robustly acceptable sentences of the language. Now we have our three familiar possibilities: inconsistency, incompleteness, and inexistence.

Field's preferred option (expressed most explicitly in discussion) is, in fact, Horn 3, inexistence. The notion makes no sense. At first blush, this looks like a straightforward self-refutation. Hasn't Field shown how to define the predicate in set-theoretic terms? No. What he has done is shown how, given any interpretation of *ZF*, $\mathcal{M}$, to define the predicate $Val_{\mathcal{M}}(x) = 1$, '*x* has value 1 relative to interpretation $\mathcal{M}$'. It is the absolute notion, which is not explicitly defined in the construction (nor can it be, or we would have a consistency proof for the theory, and so for *ZF*, in *ZF*), the existence of which Field denies.

This notion would appear to be presupposed by the construction in a very robust way, though. The whole point of Field's construction is to delineate and justify the inferences we are allowed to deploy in the language in question (Field's own language). It will do so only if the language has a semantic structure of the same kind as that of the interpretations that Field specifies. As someone (I forget who) said, truth in a model must be a model of truth. Having value 1 in a model must be a model of having value 1. If there is no such notion, then the fact that $Val_{\mathcal{M}}$ can be deployed to give a certain notion of validity provides *no reason whatsoever* to suppose that the notion applies to Field's language. Clearly, Field thinks it does, since he often appeals to the notion of validity he delineates to justify the legitimacy or otherwise of forms of reasoning in the language he uses. If he denies the existence of this notion, he cannot claim that certain ways of reasoning in the language he uses are legitimate—or illegitimate. Since he does make such claims, and takes his semantics to justify these, he does presuppose the notion.

There is more to be said on this matter, but before we turn to this, let us look briefly at the other two horns, inconsistency and incompleteness. The first of these is that *Val* is expressible in the language. If it is expressible in the purely set-theoretic language, then we have, for any *A*, $Val(\langle A \rangle) = 1 \vee Val(\langle A \rangle) \neq 1$, and the familiar classical reasoning gives a contradiction. If *Val* is definable in the language, but not the purely set-theoretic language, then we need not have all instances of the Law of Excluded Middle for $Val(\langle A \rangle) = 1$, and the argument to paradox is blocked. But we may now legitimately ask how to define it. We cannot define it using Field's *D* operator. Even within an interpretation, $\mathcal{M}$, is not the case that this applies to all and only the things that have value 1 in $\mathcal{M}$ (and the same goes for all the other operators in Field's *D* hierarchy). But even given a definition, we have a dilemma. If $Val(\langle A \rangle) = 1 \vee Val(\langle A \rangle) \neq 1$, we have the paradox with us; if not, our previous

worry is exacerbated. For any $\mathcal{M}$, $Val_{\mathcal{M}}(\langle A \rangle) = 1 \vee Val_{\mathcal{M}}(\langle A \rangle) \neq 1$; so having value 1 in a model is *not* a good model of having value 1.

The final possibility is the incomplete case: *Val* is a meaningful notion, but not expressible in the language. In that case, the solution fails for the standard reason: the construction has not shown potentially inconsistency-generating semantic notions to be inconsistent. It should be noted that the Definition of *Val* can be carried out in second-order *ZF*—not the second-order version of Field's theory; just second-order *ZF*. We can simply apply Field's construction to the model of first-order *ZF* that second-order *ZF* gives us.[4] Hence, Field must deny the legitimacy of second-order *ZF*, a point which considerably ratchets up the stakes in the inexistence horn.

One way or another, then, Field is subject to the revenge syndrome.[5]

9.5 Revenge and ZF

The problem with Field's preferred approach, as we have seen, is that he denies the existence of a notion that, in all honesty, he has to presuppose. Note, however, that this is not a problem of Field's theory as such. It is simply one that the theory inherits from orthodox set-theory, *ZF*.

Suppose, *per*, one might hope, *impossibile*, that one could define the intended interpretation of the language of *ZF* in *ZF*. Then Field's construction would give us an intended interpretation for his extended language, and his theory of validity would be applicable to its own language. The failure to be able to do this is, then, simply a result of a certain inability of *ZF*.

Since *ZF* is normally taken to be perfectly kosher, this might suggest that there is nothing to worry about. But there is. If one were to attempt to specify the intended interpretation for *ZF* in *ZF*, it would have the structure $\langle V, \in_V \rangle$, where V is the set of all sets (or, assuming the Axiom of Foundation, the Cumulative Hierarchy), and $\in_V$ is the membership relation on V. But one cannot do this, since V is not a set. Now, what are we to say of this V? There are three possibilities.

1. Inconsistency. The first is that the existence of V can be recognized in *ZF*. In some way, we can prove its existence. In this case, of course, *ZF* would be inconsistent. We would be able to prove the consistency of *ZF* in *ZF*, and so Gödel's second incompleteness theorem would kick in. This is like using a sledge hammer to crack a nut, however. Since *ZF* entails the non-existence of V, we have an immediate contradiction.

[4] See Rayo (2007).

[5] This is not the only problem with Field's account. For a discussion of revenge and other problems, see Priest (2006).

2. Incompleteness. The second is to suppose that V exists, but that it is not one of the sets in *ZF*. But this would show that *ZF* is not the theory of *all* collections, which it was supposed to be. Nor does it help to suppose that V is a proper class, assuming this notion to make sense. Proper classes would seem to be just the next layer up in the cumulative hierarchy. If the sets of *ZF* really exhaust the hierarchy, there is no next layer. Even if there were, exactly the same problem would then arise with respect to the totality of all classes (proper and otherwise). So the problem has not been solved; merely relocated.

3. Inexistence. For that reason, the only really robust possibility for a solution is to deny the existence of V *tout court*. But that's problematic. Reasoning in natural ways, we would appear to make legitimate use of large totalities such as V on many occasions. The tendency to invoke proper classes is but a manifestation of this fact. For example, the natural understanding of various categories in category theory, such as the category of all sets (let alone the category of all categories) is exactly about such totalities.

And *ZF* itself would appear to *presuppose* the existence of V. For a start, the model-theoretic account of validity given in *ZF* cannot be applied to the language of *ZF* itself unless $\langle V, \in_V \rangle$ is an interpretation, which it is not. In other words, just as for Field, the logic the theory defines is not applicable to the theory itself, and we are bereft of a justification for reasoning about sets, one way or the other.[6]

Other considerations point in the same direction. A quantified sentence has no determinate truth-value unless the range of the quantifiers is a determinate totality. If I say 'everyone has the right to vote', what I say is true if restricted to adults, false if it includes minors. But in *ZF* we quantify over all sets, and we take it that the theorems of *ZF* are determinately true. There must, then, be a determinate totality of all sets. (Call this a proper class if you want. The name is unimportant.) Just as in the case of the *Tractatus*, denying the existence of V is therefore tantamount to denying that statements of *ZF* have meaning—or at least determinate meaning. For a second reason, then, *ZF* seems to presuppose a totality the existence of which it denies.[7]

[6] The problem is well recognized by classical logicians. One solution is proposed by Kreisel (1967). By appealing to a pre-theoretic notion of validity and its supposed properties, he argues that we may take the absolute notion of validity to be extensionally equivalent to the model-theoretic notion. One might have various objections to Kreisel's argument; but in any case, the strategy is unlikely to appeal to Field, just because he, unlike Kreisel, is trying to drive a wedge between the model-theoretic situation and the absolute situation. In particular, the absolute notion of validity, presupposing as it does the function *Val*, must also, according to him, be meaningless.

[7] Further on all these matters, see Priest (1987), ch. 2, and (1997), ch. 11. The problem discussed in this section is essentially that generated by Cantor's paradox of the greatest cardinal size. But, as one might expect, Burali–Forti's paradox of the greatest ordinal size produces essentially the same problem. See Shapiro (2007).

The revenge problem is normally thought of as applying to the semantic paradoxes, not the set-theoretic paradoxes, but as we have just seen, the revenge situation is exactly the same for set theory, at least for *ZF*.[8] The set *V* is not itself a semantic notion, but it is conceptually closely connected with the semantics of *ZF*.[9] And one has the same three options: inconsistency, incompleteness, and inexistence—each with its own come-uppance.

Any solution to the semantic paradoxes which piggybacks on *ZF*, such as Field's, whatever it says about truth, is, therefore, subject to revenge problems. The revenge of *V*.

References

Beall, JC, and Armour-Garb, B. (2006). *Deflationism and Paradox*, Oxford: Oxford University Press

Field, H. (2003). 'A revenge-immune solution to the semantic paradoxes', *Journal of Philosophical Logic* 32, 139–77

——— (2006). 'Is the liar both true and false?', ch. 2 of Beall and Armour Garb (2006)

——— (2007). 'Solving the paradoxes, escaping revenge', this volume

Kreisel, G. (1967). 'Informal rigour and completeness proofs', pp. 138–71 of I. Lakatos (ed.), *Problems in the Philosophy of Mathematics*, Amsterdam: North Holland

Priest, G. (1987). *In Contradiction*, Dordrecht: Martinus Nijhoff. 2nd edn. Oxford: Oxford University Press (2006)

——— (1995). *Beyond the Limits of Thought*, Cambridge: Cambridge University Press. 2nd edn. Oxford: Oxford University Press (2002)

——— (2006). 'Spiking the Field artillery', ch. 3 of Beall and Armour Garb (2003)

Rayo, A. and Welch, P. D. (2007). 'Field on revenge', this volume

Shapiro, S. (2007). 'Burali–Forti's revenge', this volume

[8] There are some non-standard set theories, such as Quine's *NF*, which have a universal set. Obviously, the considerations I have applied to *ZF* do not carry over immediately to them. However, there will be similar considerations in such cases. For example, although such theories may well be able to define their own intended interpretation, what, then, they cannot do, on pain of inconsistency, is prove that the theory holds in that interpretation. Again, we are bereft of a justification for supposing that the theory applies.

[9] See Priest (1987), 2.5.

10

Field on Revenge

Agustín Rayo and P. D. Welch

In a series of recent papers,[1] Hartry Field has proposed a novel class of solutions to the semantic paradoxes, and argued that the new solutions are 'revenge-immune'. He has argued, in particular, that by building on a sufficiently expressive language one can get a language which is able to express its own semantic theory, including its own truth predicates and any intelligible determinacy predicates. The purpose of this note is to argue that the plausibility of Field's revenge-immunity claim depends crucially on the status of higher-order languages. We show that by availing oneself of higher-order resources one can give an explicit characterization of the key semantic notion underlying Field's proposal, and note that inconsistency would ensue if the languages under discussion were expressive enough to capture this notion.

10.1 Field's Proposal

For the sake of concreteness, we shall consider the version of the proposal developed in Field (2003a) and further elucidated in Field's contribution to the present volume. Start with a standard first-order language L containing the set-theoretic primitives '$\in$' and 'Set(. . .)', together with an arbitrary selection of non-set-theoretic predicates.

Thanks to Hartry Field, Hannes Leitgeb, and Graham Priest for their many helpful comments. Special thanks are due to Vann McGee. PDW acknowledges support from the British Academy and EPSRC Research Grant (EP/C531485/1).

[1] Field (2003a), Field (2003b), Field (2004), and Field's contribution to the present volume.

On its intended interpretation, the range of L's quantifiers includes all sets, '$\in$' expresses set-theoretic membership and 'Set(. . .)' is true of all and only sets. In the standard, set-theoretic sense of 'model', there is no model that captures L's intended interpretation, since models are sets and, in standard set theory, no set has a transitive closure including all sets. But Field invites us to consider the next best thing to an intended model for L: the *quasi-correct* models of L. A quasi-correct model for L is a model of L in which '$\in$' expresses set-theoretic membership amongst individuals in the extension of 'Set(. . .)', and the extension of 'Set(. . .)' consists of all sets built up from the model's urelements up to some inaccessible rank (where the model's urelements are those of the individuals in the domain of the model that are not in the extension of 'Set(. . .)').

So far, nothing out of the ordinary has happened. The action comes when Field extends L to a richer language L^+ containing the new one-place predicate 'Tr(. . .)' and the new two-place sentential connective '$\longrightarrow$', and uses an iterated version of the construction in Kripke (1975) to characterize what might be called the *completion* of a quasi-correct model m of L. We needn't concern ourselves with the details of Field's construction for now, except to note that if m^+ is the completion of a quasi-correct model m of L, then:

1. m^+ is a many-valued model for L^+, in which every sentence gets assigned value 1, 0, or $\frac{1}{2}$;
2. every sentence of L that is true in m gets value 1 in m^+, and every sentence of L that is false in m gets value 0 in m^+;
3. every instance of the truth-schema '$\mathrm{Tr}(\langle A \rangle) \longleftrightarrow A$' gets value 1 in m^+ (where $\ulcorner \phi \longleftrightarrow \psi \urcorner$ is an abbreviation for $\ulcorner (\phi \longrightarrow \psi) \wedge (\psi \longrightarrow \phi) \urcorner$); and
4. intersubstitution of A and $\ulcorner \mathrm{Tr}(\langle A \rangle) \urcorner$ in an arbitrary sentence of L^+ does not change the value it gets assigned by m^+.

Moreover, when validity in L^+ is defined as preservation of value 1 under all uniform substitutions of non-logical vocabulary in the completion of a quasi-correct model of L, one gets a weakening of classical logic that Field calls LCC (short for 'the Logic of Circularly Defined Concepts'). LCC fails to validate excluded middle or the equivalence between $\ulcorner \phi \longrightarrow \psi \urcorner$ and $\ulcorner \neg \phi \vee \psi \urcorner$, but is strong enough to be interesting and has a number of attractive features. And an immediate consequence of clause 2 above is that LCC behaves classically for the special case of formulas not containing 'Tr(. . .)' or '$\longrightarrow$'.

These results put Field in a position to explain how one might acquire a concept of truth for L^+ that would allow one to accept every instance of the truth-schema. One starts out speaking L and reasoning classically, and sets out to speak L^+. The first step is to acquire a novel understanding of the logical vocabulary:

One makes the assumption that the logical vocabulary in L^+ is to be understood on the basis of its role in LCC-inference, and uses the concept of LCC-validity to develop such an understanding.

Since LCC-inference behaves classically when restricted to formulas of L there is no risk that one's novel understanding of the standard logical vocabulary will force one to give up inferences one previously endorsed. And since the notion of LCC-validity is characterized within standard-set theory, which is expressible in L, there is no risk of circularity.

The next step is to acquire an understanding of the new predicate 'Tr(...)'. This can be done as follows:

One resolves to accept every L^+-instance of '$\mathrm{Tr}(\langle A \rangle) \longleftrightarrow A$' , and makes the assumption that 'Tr(...)' is to be understood on the basis of its role in this schema.

But how does one know that this won't lead to inconsistency? It is here that Field's *pièce de résistance* comes in: one can use the observation that any quasi-correct model of L has a completion to show that the result of enriching ZFC with every instance of the truth-schema has no untoward LCC-consequences. And since, given an arbitrary quasi-correct model, Field's construction can be carried out entirely within standard set-theory, it can be carried out using the expressive resources of L and therefore the expressive resources of a classically behaving fragment of L^+. So one can convince oneself that ZFC plus every instance of the truth-schema has no untoward LCC-consequences without ascending to a strictly richer metalanguage. (On the other hand, one can only prove that L has a quasi-correct model by assuming a theory at least as strong as ZFC plus a large-cardinal hypothesis. So one's warrant for the claim that one's preferred system of set-theory plus every instance of the truth-schema is LCC-consistent presupposes a warrant for the claim that a suitable large-cardinal hypothesis is true.)

10.2 A Small Interlude

Field's *pièce de résistance* shows that ZFC plus every instance of the truth-schema has no untoward LCC-consequences. But, as Vann McGee pointed out to one of us in conversation, this does not immediately guarantee that one won't be able to derive a contradiction in LCC from true sentences of L and instances of the truth-schema. For it is consistent with ZFC that no quasi-correct model of L assigns every sentence in L its intended truth-value. And if there is no such model, no completion of a quasi-correct model is such that every true sentence of L and every L^+-instance of the truth-schema is assigned value 1.

Here is an analogy. Suppose there are precisely thirteen inaccessibles. Then if I_{13} is a sentence of L stating that there are thirteen inaccessibles, I_{13} is false according to any quasi-correct model of L, but true *simpliciter*. So although the existence of quasi-correct models of L guarantees that ZFC plus the negation of I_{13} is consistent, it does not guarantee that one won't be able able to derive a contradiction from true sentences of L and the negation of I_{13}.

Field's *pièce de résistance* makes it seem extremely plausible that one won't be able to derive a contradiction in LCC from true sentences of L and instances of the truth-schema. But it is important to be clear that all Field has given us is a plausibility argument. An immediate consequence of the result in section 5 is that one can transform Field's plausibility argument into a proof by availing oneself of higher-order resources. The result also entails that Field's plausibility argument can be transformed into a proof by assuming the existence of a κ such that V_κ is a Σ^1_3-substructure of V (given a suitable definition for the notion of a Σ^1_3-substructure).

10.3 Revenge

Field has a plausible argument for the claim that the truth-schema won't lead to inconsistency in LCC. We would like to suggest, however, that he only gets consistency because the language under discussion is unable to express a key semantic notion. We will see that the notion of an intended interpretation for L^+ can be characterized using higher order resources, and that inconsistency would immediately ensue if the intended interpretation of L^+ was expressible in L^+.

Consider a warm-up case. Suppose we took our initial object language to be the language of first-order arithmetic, L_A, rather than L. Whereas there is no model corresponding to the intended interpretation of L, it is easy to construct a model, m_A, that captures the intended interpretation of L_A: m_A is just the standard model of arithmetic, $\mathbb{N}$. Field's construction can be applied to m_A just as it was applied to quasi-correct models of L. What one gets is a many-valued model m_A^+ for L_A^+ (which is the result of extending L_A with 'Tr(. . .)' and '$\longrightarrow$'). But since m_A is the intended model of L_A, one can expect m_A^+ to count as the intended model of L_A^+. In other words, one can expect the value m_A^+ assigns a sentence of L_A^+ to be its 'real' truth-status. One might say, for example, that a sentence of L_A^+ ought to be *accepted* just in case it is assigned the value 1 by m_A^+, and *rejected* just in case it is assigned value 0 or $\frac{1}{2}$. It is a consequence of Field's construction that the value of a negation is always 1 minus the value of its negatum. So whereas sentences of L_A^+ whose m_A^+-value is $\frac{1}{2}$ are such that both they and their negations ought to be rejected, sentences of L_A^+ whose m_A^+-value

is 0 are such that their negations ought to be accepted even though they themselves ought to be rejected.

This yields all the right results. Every arithmetical truth, every instance of the truth-schema for L_A^+ and all of their LCC-consequences are assigned m_A^+-value 1, so they ought to be accepted and their negations ought to be rejected. The Liar sentence Q is assigned m_A^+-value $\frac{1}{2}$, as are $\ulcorner\neg Q\urcorner$, $\ulcorner \mathrm{Tr}(\langle Q\rangle)\urcorner$, $\ulcorner\neg \mathrm{Tr}(\langle Q\rangle)\urcorner$, and the corresponding instances of excluded middle: $\ulcorner Q \vee \neg Q\urcorner$ and $\ulcorner \mathrm{Tr}(\langle Q\rangle) \vee \neg \mathrm{Tr}(\langle Q\rangle)\urcorner$. So one can say—as Field does in discussing the set-theoretic case—that 'one must reject the claim that $\mathrm{Tr}(\langle Q\rangle)$ and also reject the claim that $\neg\mathrm{Tr}(\langle Q\rangle)$ [. . .] We must likewise reject the corresponding instance of excluded middle'. (Chapter 4, p. 89). And, of course, the *pièce de résistance*: one can use m_A^+ to show that sentences that ought to be accepted will never have untoward LCC-consequences. (One can't carry out the proof in L_A, since m_A^+ can't be characterized in L_A: it is a consequence of Tarski's Theorem that there is no formula $\phi(x)$ of L_A such that a sentence S of L_A^+ has m_A^+-value 1 just in case $\phi\langle S\rangle$ is true.)

An account of truth for the language of arithmetic based on this idea would have many virtues. But it should be clear that revenge-immunity is not one of them. Say that a language is *revenge-immune* just in case it is able to express its own semantic theory, including its own truth predicate and any intelligible determinacy predicates. Then it should be clear that L_A^+ is not revenge-immune. For inconsistency would immediately ensue if L_A^+ were able to express the notion of acceptability—or, equivalently, the notion of having m_A^+-value 1.[2] So although it is true that the result of adding every L_A^+-instance of the truth-schema to the standard axioms of arithmetic is LCC-consistent, this is merely because L_A^+ is unable to express the notion of having m_A^+-value 1.

So why couldn't one generate an analogous revenge problem for the set-theoretic case? The crucial observation is that the argument in the arithmetical case makes use of the fact that m_A is an intended model for the language of arithmetic, and, as emphasized above, the language of set-theory has no intended models. (Had m_A been chosen to be an unintended model for L_A, there would have been no reason for thinking that its completion should count as an intended model for L_A^+, and the proof in footnote 2 would have been open to a charge of equivocation by failing to distinguish between being assigned value 1 by m_A^+ and having 'real' value 1.)

[2] *Proof:* Working within a classical metalanguage, assume there is an open formula '$One(x)$' of L_A^+ such that $\ulcorner One(\langle A\rangle)\urcorner$ gets m_A^+-value 1 just in case A gets m_A^+-value 1 and use Gödelian techniques to find a sentence Q^* of L_A^+ such that $\ulcorner Q^* \longleftrightarrow \neg One(\langle Q^*\rangle)\urcorner$ gets m_A^+-value 1. The fact that $\ulcorner Q^* \longleftrightarrow \neg One(\langle Q^*\rangle)\urcorner$ gets m_A^+-value 1 guarantees that Q^* gets m_A^+-value 1 just in case $\ulcorner\neg One(\langle Q^*\rangle)\urcorner$ gets m_A^+-value 1. But the way in which '$One(x)$' was introduced guarantees that $\ulcorner\neg One(\langle Q^*\rangle)\urcorner$ gets m_A^+-value 1 just in case Q^* fails to get m_A^+-value 1.

What would happen if one tried to apply a version of the revenge argument to the set-theoretic case on the basis of a *quasi-correct* model of L, rather than an intended model? Suppose one has defined in L a quasi-correct model m of L, and let m^+ be the completion of m. The notion of having m^+-value 1 can be defined within L, since L includes the language of first-order set-theory. So the notion of having m^+-value 1 is expressible in L^+. But, as Field is careful to point out, m doesn't assign every sentence of L its intended truth-value (since m is definable in L). So thinking that m^+ generates a revenge problem for L^+ would be like thinking that one violates Tarski's Theorem by characterizing in L the notion of truth-according-to-m. Tarski's Theorem is not violated because truth-according-to-m is not genuine truth; similarly, revenge problems are averted because having m^+-value 1 does not correspond to genuine acceptability.

If ZFC + 'there is a Mahlo Cardinal' is consistent, then so is ZFC + 'there is an inaccessible κ such that V_κ is an elementary substructure of V'. It is a consequence of Tarski's Theorem that such a κ cannot be defined in L. But if V_κ exists nonetheless, then there is a quasi-correct model m_κ of L (not definable in L) in which every sentence of L receives its intended truth-value. Moreover, Field's construction guarantees the existence of a completion m_κ^+ of m_κ (also not definable in L) in which every sentence of L receives its intended truth-value. The existence of m_κ^+ need not generate a revenge problem for L^+, since there is no guarantee that m_κ^+ will assign every sentence of L^+ its 'real' truth-status. But it is a consequence of the construction in section 5 that if V_κ is a Σ_3^1-elementary substructure of V, then m_κ^+ assigns every sentence of L^+ its 'real' truth-status. And this *would* generate revenge problems for L^+, since inconsistency would immediately ensue if L^+ was able to express the notion of being assigned value 1 by m_κ^+.

In the following section we will see that, when higher-order resources are brought on board, there is an argument against Field's revenge-immunity claim that does not depend on the existence of such a κ.

10.4 SO-Models

By availing oneself of higher-order resources, it is possible to characterize the intended interpretation of L. Details are supplied in Rayo and Uzquiano (1999), but the basic idea is straightforward. In standard model-theory one says of an individual—a set with certain properties—that it is a model. But by availing oneself of higher-order resources one could also say of some individuals—some ordered-pairs with certain properties—that they (jointly) form an SO-model (short for 'second-order model'). If the Gs form an SO-model, one says that x is in the 'domain' of the Gs just in case

⟨'∀', x⟩ is one of the Gs, and one says that the atomic predicate-letter P is true of x according to the Gs just in case ⟨P, x⟩ is one of the Gs. Truth and satisfaction can then be characterized along familiar lines. For any standard model m there are some things that form an SO-model with the same domain as m and the same interpretations for atomic predicates as m (and therefore the same truths as m). But the big advantage of SO-models is that they allow for domains too big to form a set. There are, in particular, some individuals that together form the intended SO-model for L. Let us call them 'the Ms'. (Something is one of the Ms just in case it is either an ordered-pair of the form ⟨'∀', x⟩ (for x an arbitrary object), an ordered-pair of the form ⟨'Set', α⟩ (for α a set), an ordered-pair of the form ⟨' $\in$ ', ⟨α, β⟩⟩ (for α and β sets such that $\alpha \in \beta$), or an ordered-pair of the form ⟨$\ulcorner P_i^n \urcorner$, ⟨$x_1$, . . . , x_n⟩⟩ (for $\ulcorner P_i^n \urcorner$ true of x_1, . . . , x_n)).

In order to generate revenge problems for Field, it is not enough to show that the Ms capture the intended interpretation of L, one must also find a way of extending the Ms to an intended interpretation of L^+ by applying an analogue of Field's construction. But, as we shall see in section 5, Field's construction can be emulated in a higher-order setting. Let the M$^+$s be the result of applying the higher-order version of Field's construction to the Ms. The M$^+$s form a many-valued SO-model for L^+. And since the Ms are the intended model of L, one can expect the M$^+$s to count as the intended model for L^+. In particular, one can expect the value that the M$^+$s assign a sentence of L^+ to be its 'real' truth-status. As in the arithmetical case, one might say that a sentence of L^+ ought to be *accepted* just in case it is assigned the value 1 by the M$^+$s, and *rejected* just in case it is assigned value 0 or $\frac{1}{2}$. It will still be a consequence of the construction that the value of a negation is always 1 minus the value of its negatum. So whereas sentences of L^+ that get assigned the value $\frac{1}{2}$ by the M$^+$s are such that both they and their negations ought to be rejected, sentences of L^+ that get assigned the value 0 by the M$^+$s are such that their negations ought to be accepted even though they themselves ought to be rejected.

As in the arithmetical case, this yields all the right results. Every truth of L, every instance of the truth-schema for L^+ and all of their LCC-consequences are assigned value 1 by the M$^+$s, and therefore ought to be accepted and have negations that ought to be rejected. The Liar sentence Q is assigned value $\frac{1}{2}$ by the M$^+$s, as are $\ulcorner \neg Q \urcorner$, $\ulcorner \mathrm{Tr}(\langle Q \rangle) \urcorner$, $\ulcorner \neg \mathrm{Tr}(\langle Q \rangle) \urcorner$, and the corresponding instances of excluded middle: $\ulcorner Q \vee \neg Q \urcorner$ and $\ulcorner \mathrm{Tr}(\langle Q \rangle) \vee \neg \mathrm{Tr}(\langle Q \rangle) \urcorner$. So one can say, with Field, that 'one must reject the claim that Tr(⟨Q⟩) and also reject the claim that ¬(Tr⟨Q⟩) [. . .] We must likewise reject the corresponding instance of excluded middle'. Finally, one can use the M$^+$s to show that sentences that ought to be accepted will never have untoward LCC-consequences.

As in the arithmetical case, the resulting picture has many virtues. But revenge-immunity is not one of them. Inconsistency would immediately ensue if L^+ were able to express the notion of acceptability—or, equivalently, the notion of being

assigned value 1 by the M^+s. (The proof is exactly analogous to that in footnote 2.) So although it is true that the result of adding every L^+-instance of the truth-schema to the standard axioms of set theory is LCC-consistent, this is merely because of L^+s inability to express the notion of being assigned value 1 by the M^+s.[3]

Needless to say, the argument against revenge-immunity in the set-theoretic setting will be a non-starter unless one is willing to countenance the legitimacy of classical higher-order metalanguage. The status of higher-order languages is the subject of intense debate. We cannot hope to address this debate here.[4] All we wish to show is that the plausibility of Field's revenge-immunity claim is sensitive to the outcome.

10.5 The Construction

In order to generate revenge problems for Field, it is not enough to show that the Ms capture the intended interpretation of L, one must also find a way of extending the Ms to an intended interpretation of L^+ by applying an analogue of Field's construction. In order to do this we sketch the development of a fragment of second order set theory over V, the class of all sets. (Talk of classes is a notational convenience: first-order quantification over classes is to be thought of as a syntactic abbreviation for second-order quantification.[5])

At the heart of Field's construction over a suitable first-order model m is a classical recursive process along an initial segment of the ordinals. The clauses of this recursion interleave a hybrid constuction of obtaining a Kripkean fixed point using the strong Kleene scheme of partial logic with an evaluative process for $\longrightarrow$ which encapsulates a history of what has happened at previous stages. At each ordinal stage, (i) all sentences involving Tr are reset to have value $\frac{1}{2}$; (ii) the semantic value of sentences involving $\longrightarrow$ are calculated according to the aforementioned process; (iii) finally a new fixed point à *la* Kripke is computed. The latter process assigns values to sentences involving Tr (but does not alter the values of the binary operator $\longrightarrow$; these remain

[3] For further discussion of this point, see Priest's contribution to this volume.

[4] For the conservative side of the debate see Quine (1986), ch. 5, Resnik (1988), Parsons (1990), and Linnebo (2003), among others. For the liberal side of the debate see Boolos (1984), Boolos (1985a), Boolos (1985b), McGee (1997), Hossack (2000), McGee (2000), Oliver and Smiley (2001), Rayo and Yablo (2001), Rayo (2002), Williamson (2003), and Rayo (forthcoming), among others.

[5] A note about our use of higher-order resources: the upshot of our proof is that a certain version of second-order set-theory with restricted second-order comprehension is enough to prove that the M's can be extended to an intended interpretation of L^+. Accordingly, the existence of an intended interpretation of L^+ is a purely second-order result. In the course of proving this result we sometimes indulge in talk of collections of classes, but this is for expositional convenience only: the proof can be carried out entirely within a second-order language.

fixed through the process of calculating the next fixed point). At the beginning of each ordinal stage α say, during (ii), essentially a Σ_2-recursive clause is invoked: the semantic value of $A \longrightarrow B$: the value here depends on whether *there exists* a previous stage β so that *for all* subsequent stages γ before α something happens about the valuations of A and B at those stages γ.

It is this latter clause that gives the construction its essential flavour, and its overall complexity. Nevertheless, as m is a set, there will be some least *acceptable ordinal* $\Delta_0(m)$ at which the whole process stabilises out and starts to cycle. Field (2003a) demonstrates the existence of such acceptable ordinals, and in Welch (2003) several equivalent characterisations of such are given. Essentially Welch (2003) shows that these characterisations can be obtained by a demonstration that over $m = \mathbb{N}$, the standard model of arithmetic, the set of sentences that have 'ultimate' semantic value 1 (i.e. will receive semantic value 1 at some point never to be later changed) is recursively isomorphic to Herzberger's set of *'stable truths'* obtained from a single revision sequence starting from a distribution of semantic value 0 to all sentences. Other characterisations of this set were already known, and yielded a computation of Field's least accetptable ordinal $\Delta_0(\mathbb{N})$.

That article also mentioned that the whole construction over $\mathbb{N}$ could be performed in second-order number theory (the 'Second Demonstration'); although it urged the reader not to do so (since the possibility of doing so was given explicity by the argument above—the 'Third Demonstration' and the details would be wearisome) it is in fact this 'second demonstration' that we must perform here, albeit translated to the second order set-theoretical arena, rather than the number-theoretic one.

As V contains all ordinals we must elaborate a theory of *wellorderings* given by class terms that are of sufficient length to allow us to prove that Field's recursive construction can be emulated along these *wellorderings*. In case the reader is a little queasy about the notion of such orderings, we could point out that it is easy in standard first order ZF set theory to define orderings that have the apparent 'order type' 'On + On' where On denotes the class of all ordinals: one simply defines a class of pairs $\langle \alpha, i \rangle$ for $\alpha \in \mathrm{On}, i \in \{0, 1\}$ and defines $\langle \alpha, i \rangle <' \langle \beta, j \rangle \longleftrightarrow i < j \vee (i = j \wedge \alpha < \beta))$ (where $<$ has its usual meaning). It is simply that such order-types are not *set-like*, that is they do not have initial segments that are sets. Nevertheless with care, set theorists can, and do, use such orderings. Similar definitions are immediate for $\mathrm{On} \times \mathrm{On}$, $\mathrm{On}^{\mathrm{On}}$ etc., etc. However the ordering type required in Field's construction cannot be given by any first-order class terms definable over V. Thus, for example, the construction cannot be done in vNGB set theory. Indeed we shall see that instances of Π^1_3 Comprehension in second-order set theory are needed (which we shall call Π^1_3−CA by analogy with that of subsytems of analysis (see (Simpson 1999))).

In the following L will be the standard countable first-order language for set theory with equality and set membership symbol '$\in$' as the sole binary predicates, and without

constants. We shall add constants to L, one for each set $x \in V$. Developing the syntax for this language using sets as codes can be done in a weak set theory (and so in ZFC) (see, for example Devlin 1984, ch. I. 9 for a standard account). We shall call this language L_V; done over over $(V, \in)$, this yields a class term for the elements of this language: thus $L_V \subseteq V$.

These languages augmented with the predicate symbol Tr and a binary conditional relation $\longrightarrow$ will be called L^+, and L_V^+ respectively (the latter language is also then given by a simply defined class term over $(V, \in)$).

We sketch the adaptation of the first stage of Field's construction (now over $(V, \in)$). This first starts out with assigning semantic values of $\frac{1}{2}$ to all atomic sentences of L_V^+ of the form $\mathrm{Tr}(u)$ where u is (a set coding) a sentence of L_V. Similarly all sentences of the form $A \longrightarrow B$ where $A, B \in L_V$ are also set to value $\frac{1}{2}$. The rest of the clauses for this first stage are those needed to construct in a familiar manner the first strong Kleene fixed point over the structure $(V, \in)$ in the language L_V^+. Those sentences assigned a semantic value of 1 then form a proper subclass of L_V^+. This class is not given by a term *definable* over $(V, \in)$ but is the result of an *inductive* second order process over $(V, \in)$. (Similarly for the classes of sentences with semantic values 0, $\frac{1}{2}$.) The use of the language L_V^+ enriched with constants for all sets allows a simple substitutional quantification clause for the working out the values of $|\forall \upsilon A(\upsilon)|_{0,\sigma}$ and $|\exists \upsilon A(\upsilon)|_{0,\sigma}$ (to use Field's notation for the semantic values calculated at the σth substage of this—the first full stage of the cumulative process of computing Strong Kleene fixed points.)

Supposing that (X^1, X^0) is a pair of classes contained in L_V^+ with X^1 those receiving value 1, X^0 receiving value 0 (and the rest by default $\frac{1}{2}$), at some substage. We can view one step of this process as a (Strong Kleenian) *jump operation*: $j((X^1, X^0)) = (Y^1, Y^0)$ which delivers the extensions of those classes receiving the respective semantic values at the next stage: $u \in Y^1$ (and so in the extension of Tr) at the next substage, if at the previous substage certain 'truth conditional' clauses are fulfilled. (We'll set $j^1((X^1, X^{-0})) = Y^1$ and $j^0((X^1, X^0)) = Y^0$.) For any particular u these are elementary conditions on $(V, \in, (X^1, X^0))$. Overall we may say that the relations

$$u \in j^1((X^1, X^0)); u \in j^0((X^1, X^0))$$

whilst not elementary for disjoint classes $X^1 \cap X^0 = \varnothing$, are Δ_1^1 over $(V, \in)$. (A Π_1^1 relation $\mathcal{R}(\upsilon, Z)$ over $(V, \in)$ is one of the form $\mathcal{R}(\upsilon, Z) \Longleftrightarrow \forall U \varphi(\upsilon, Z, U)$ where φ is elementary in the language L_V with additional second order variables Z, U. A Σ_1^1 relation is the complement of a Π_1^1 relation, and a relation is Δ_1^1 if both it and its complement can be written in Π_1^1 form. For the case in hand the relations are not quite elementary since u could be the code of a sentence of arbitrary complexity in the usual Levy hierarchy of classification of formulae, but the relations are not far

from elementary.) In Field's notation, if (X^1, X^0) are the classes of sentences assigned semantic value 0/1 at stage σ then $|\mathrm{Tr}(u)|_{\sigma+1} = 1 \iff u \in j^1((X^1, X^0))$, etc.

In short this first stage of establishing the Strong Kleene fixed point in L_V^+ is a *monotone inductive* process *over* $(V, \in)$. In particular we can think of this couched in terms of the extension of the theory of inductive definability to a theory of inductive *second order* relations as adumbrated in Moschovakis (1974 ch. 6).

Similarly the second order relation $\mathcal{U}(X, Y)$:

$$X \cap Y = \varnothing \wedge j^1(X, Y) \subseteq X \wedge j^0(X, Y) \subseteq Y$$

is also Δ_1^1.

Consequently the first fixed point (A_0^1, A_0^0) where (in Field's notation $u \in A_0^1 \iff |u|_0 = 1$ and further $u \in A_0^1 \iff \mathrm{Tr}(u) \in A_0^1$, and so on) is Π_1^1-definable:

$$u \in A^+ \iff \forall X^1 \forall X^0 (\mathcal{U}(X^1, X^0) \implies u \in X^1).$$

We thus have that both A_0^1 and A_0^0 are both Π_1^1 and also inductive over $(V, \in)$. (There should be a word of warning here: over $\mathbb{N}$ Π_1^1 and (positive) inductive relations coincide, but $\mathbb{N}$ is a special case, and over other structures positive elementary inductive relations are Π_1^1, but the converse may fail in general, so not all results will generalize.) Of course the 'induction' that is implicit in the above would in the set-sized case have a particular length (again in Field's terminology, that $|u|_0$ appearing in the above is strictly $|u|_{0,\Omega}$ for some 'sufficiently large' ordinal Ω).

As already adverted, in our context a wellordering of sufficient length along which to run the Kripkean construction is beyond any length of a first-order definable over $(V, \in)$ wellordering $<$ of On. We further need also to define the semantic values of formulae of the form $A \longrightarrow B$, *after* each fixed point has been reached. In particular this whole process of finding fixed points and evaluating 'conditionals' will have to be repeated for more stages than there are ordinals, for a particular liminf limit rule to be applied at 'limit' stages. We thus instead resort to a weak second-order set theory.

We reserve upper case letters T, W, U, to informally denote classes (as we have been doing), with X_i, etc. to be class variables. Sets t, w, u will be lower case, with x_i etc. being set variables. An $\mathcal{L}^2$ *structure* is as follows:

$$\mathfrak{M} = (V, S_M, \in)$$

where $S_M \subseteq \mathcal{P}(V)$ is a collection of *classes*, and is used to interpret the second-order variables of $\mathcal{L}^2$. If $\mathcal{B}$ is any subclass of $V \cup S_M$ then $\mathcal{L}^2_{\mathcal{B}}$ is the language of $\mathcal{L}^2$ augmented

by constants from $\mathcal{B}$. A class $W \subseteq V$ is *definable over* $\mathfrak{M}$ *using parameters from* $\mathcal{B}$ if there exists a formula $\varphi(v_0) \in \mathcal{L}^2_{\mathcal{B}}$ so that $W = \{w \in V | \mathfrak{M} \models \varphi(w)\}$. We take as axioms:

Definition 1 Γ-Comprehension Axioms *(Γ-CA) comprise*:

(i) *The usual* ZFC *axioms for sets*;

(ii) *A second-order induction axiom:*

$$\forall\alpha(\forall\beta < \alpha(\beta \in X) \longrightarrow \alpha\in$$

(iii) Γ*-Comprehension scheme (where* $\Gamma \subseteq \mathcal{L}^2$ *is a class of second-order formulae*)

$$\exists X\forall y(y \in X \longleftrightarrow \varphi(y))$$

for any $\varphi \in \Gamma$.

In the above Γ will usually restricted to be of the form $\mathbf{\Sigma}^1_1$, $\mathbf{\Pi}^n_1$ etc. We shall interpret the set variables as always ranging over the *standard universe* V: there will thus be no non-standard models. Continuing the second-order number theoretic analogy, we are assuming all models are 'ω-models'. Again, continuing the analogy:

Definition 2 $\mathfrak{M} = (V, S_M, \in)$ *is a* $\mathbf{\Sigma}^1_k$-correct model (of set theory) *iff whenever* φ *is a* $\mathbf{\Sigma}^1_k$*-sentence of* $\mathcal{L}^2$*, then* φ *is true if and only if* $\mathfrak{M} \models \varphi$.

Of course the 'φ is true' here has to be interpreted relative to some ambient domain of discussion containing $\mathfrak{M}$. This domain will later be taken to be a larger model $\widetilde{\mathfrak{M}} = (V, S_M, \in)$ with $S_M \supseteq S_M$. Without specifying it further at the moment, we consider the following discussion to be developed within this model.

We note that the assertion that some $W \in S_M$ for which $W \subset (\mathrm{On} \times \mathrm{On})$ is a class of ordered pairs which is a *linear ordering* is first-order (or 'elementarily') definable in $\mathfrak{M}$ (it is in '$\mathbf{\Pi}^1_0(\mathfrak{M})$'). (We merely have to assert that the class of pairs satisfies the usual requirements of transitivity, anti-symmetry, and trichotomy of a (strict) total order, which are simply expressed using universal quantifiers over On.) An assertion concerning 'wellordering' however is still just an elementary quantification. (For any $W \in S_M\mathfrak{M} \models$ 'W is a wellordering of On' can be expressed as '$\forall\tau \in \mathrm{On} W \cap \tau \times \tau$ is wellordered'.) Our models are therefore automatically correct about which classes wellorder (a subclass of) the ordinals. Thus for such models the notion of being a class wellordering is absolute. We further claim that that the definition of, for example, $\mathbf{\Sigma}^0_1$-satisfaction over second-order models $\mathfrak{N}$ can itself be given in a $\mathbf{\Sigma}^0_1$ fashion, by basically the same reasoning as for the first-order case.

The narrative now is developed parallel to that of second-order number theory, assuming as a base theory our structures $\mathfrak{M}$ model first-order, or 'elementary' comprehension ('ECA').

Note that a necessary and sufficient condition to be a Σ^1_1-model is that for any $X \in S_M$ that the complete $\Sigma^1_1(X)$ definable class over $\mathfrak{M}$ also be in S_M. Moreover:

Lemma 1 *Suppose* $\mathfrak{M} = (V, S_M, \in)$ *is a model of ECA. Then the following are equivalent*:

(i) $\mathfrak{M}$ *is a* Σ^1_1*-correct model of* Π^1_1-CA;

(ii) For any $X \in S_M$*, the complete* $\Sigma^1_1(X)$ *class is in* S_M.

Taking $\mathfrak{N} = (V, \varnothing, \in)$, we may look at the least collection of classes containing those first-order definable over $\mathfrak{N}$, and then closing under the operation of taking for any class X, the complete $\Sigma^1_1(X)$ class. Let S be the resulting collection of classes. We then have:

Lemma 2 $\mathfrak{M} = (V, S, \in)$ *is the minimum* Σ^1_1*-correctmodel of* Π^1_1-CA. *Similarly for any class* X *there is a minimum* Σ^1_1*-correct model of* Π^1_1-CA, $\mathfrak{M}_X = (V, S_{M_X}, \in)$ *with* $X \in S_{M_X}$.

Definition 3 *An* ordinal coded Σ^1_1-correct model *is a class* $W \subset \mathrm{On} \times V (\subset V)$ *which codes a* Σ^1_1*-correct model* $\mathfrak{M} = (V, S_M, \in)$ *via* $S_M = \{(W)_\alpha | \alpha \in \mathrm{On}\}$ *where* $(W)_\alpha = \{x | (\alpha, x) \in W\}$

We can then think of certain Σ^1_1-correct models $\mathfrak{M}$ as themselves just classes, if S_M is so enumerable by some class $W = W(\mathfrak{M})$.

Lemma 3 Π^1_1-CA $\Longleftrightarrow$ *For all* X *there is an ordinal coded* Σ^1_1*-correct model* $\mathfrak{M}_X$ *with* $X \in S_M$.

We shall need certain second-order schemes of recursion available; let us call these *class recursions*: the idea is that we suppose we have second-order formula $\varphi(x, X) \in \mathcal{L}^2$ (with possibly other set or class parameters), then given a wellordering $W \subset V \times V$, for each $x \in \mathrm{Field}(W)$ we shall associate a class Y_x by a recursion along W: if Y_z has already been defined for $z <_W x$ then $Y^x =_{\mathrm{df}} \{(u, v) | u \in Y_v \wedge v <_W x\}$, and then $Y_x =_{\mathrm{df}} \{q | \varphi(q, Y^x)\}$.

Definition 4 *The scheme of* Π^1_k-class recursion (Π^1_k *-REC) comprises all instances of the following where* $\varphi(x, W) \in \Pi^1_k$ *(possibly with other set and class parameters):*

$$\forall W(\mathrm{WO}(W) \longrightarrow \exists Y H_\varphi(W, Y))$$

where H_φ *says that* Y *is the class of all pairs* (u, x) *such that* $x \in \mathrm{Field}(W)$ *and* $\varphi(u, Y^x)$ *where* $Y^x =_{\mathrm{df}} \{(u, v) | (u, v) \in Y \wedge v <_W x\}$.

As Welch (2003) argues in the Third Demonstration, that what one needs for Field's construction over $\mathbb{N}$ to succeed is a 'Σ_2-extendible ordinal': an ordinal large enough that it enjoys a certain amount of reflection in Gödel's constructible hierarchy L built over $\mathbb{N}$. Proof-theoretically, in this Second Demonstration we need a sufficiently strong theory so that we can prove the existence of certain ordinal coded Σ^1_1-correct

models with similar sufficiently strong reflection properties: it is sufficient to have that there is at least one such Σ^1_1-correct model $\mathfrak{M}$, which enjoys Σ^1_3-reflection into an ordinal coded submodel $\mathfrak{N}$ (which perforce will be Σ^1_1-correct), which we shall write as: $\mathfrak{N} \prec_{\Sigma^1_3} \mathfrak{M}$. More formally:

Definition 5

(i) $\mathfrak{N}$ *is a* submodel *of* $\mathfrak{M}$ *if* $S_N \subseteq S_M$;

(ii) $\mathfrak{N}$ *is a* Γ-submodel *of* $\mathfrak{M}$ *if for all sentences* $\varphi \in \Gamma$ *with parameters from* $\mathfrak{N}$ *we have* $\mathfrak{N} \models \varphi \iff \mathfrak{M} \models \varphi$.

By analogy with Lemma 3 above we have at the third level:

Lemma 4 Π^1_3-CA *proves that for every* X *there is an ordinal coded* Σ^1_3*-correct model* $\mathfrak{M}_X$ *with* $X \in S_M$

We now apply Lemma 4 twice: the first time to get an ordinal coded Σ^1_3-correct model $\mathfrak{N}$ with $X \in S_N$, and then a second time to get a model $\mathfrak{M}$ with a code for $\mathfrak{N}$ in S_M. As both models are Σ^1_3-correct we obtain the right amount of reflection:

Lemma 5 Π^1_3-CA *proves that for every* X *there are ordinal coded models* $\mathfrak{N}$, $\mathfrak{M}$ *with* $\mathfrak{N} \prec_{\Sigma^1_3} \mathfrak{M}$ *and* $X \in S_N$.

If we then assume our ambient universe $M = (V, S_M, \in)$ is a model of Π^1_3-CA, then the conclusion of Lemma 5 holds. Take $X = V$ and we shall have a pair of models $(\mathfrak{N}_0, \mathfrak{M}_0)$ as in Lemma 5. Let us fix our attention on this pair.

Given a distribution of semantic values (X^1, X^0) for an L^+ model $\mathfrak{N}$, by Π^1_1-CA we can find the strong Kleene fixed point class for this distribution. To calculate the ultimate acceptable semantic values we need to iterate this process along a wellordering given to us by Π^1_3-CA, namely the wellorderings of the second-order model $\mathfrak{M}$. The successor stages are given by calculating the next Kripkean fixed point—and so are given by a Π^1_1-recursion. However at limit points of the wellordering $W = (\mathrm{Field}(W), <_W)$, we see that if say $u \in \mathrm{Field}(W)$ is a $<_W$-limit then evaluations of $|A \longrightarrow B|_u$ are given by the following

$$|A \longrightarrow B|_u = \begin{cases} 1 & \text{iff}(\exists w <_W u)(\forall v \in [w, u]_W)(|A|_v \leq |B|_v), \\ 0 & \text{iff}(\exists w <_W u)(\forall v \in [w, u]_W)(|A|_v > |B|_v) \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

(Actually the above clauses define $|A \longrightarrow B|_u$ for *any* $u \in \mathrm{Field}(W)$ but for u the successor of u_0 the evaluation is just elementary in the Kripkean fixed-point evaluations at stage u and u_0, and hence could be considered as part of an overall Π^1_1 recursion, were it not for the fact that it is at the *limit* stages where the above definition has all its

bite.) As can be seen by the further outer pair of alternating $\exists\backslash\forall$ quantifiers in this definition by cases, we shall have here a Σ^1_3-recursive clause.

We then finally have:

Lemma 6 Π^1_4-CA *proves that the Fieldian construction starting with V - the 'intended model'* of set *theory - succeeds, i.e. that there is a model* V^+.

Proof: Assume our ambient universe is a model of Π^1_4-CA: this is sufficient to iterate Π^1_3-CA along any wellorder; this means that we have Π^1_3-REC. In our case we can take any wellorder in S_{M_0} that is longer than the supremum of those in our chosen ordinal coded Σ^1_3-correct $\mathfrak{N}_0$ from Lemma 5. We thus have, in particular, if $W \in S_{M_0}$ is a wellorder of rank greater than anything in S_{N_0}, and if $u \in \text{Field}(W)$ has rank the supremum of ranks of wellorderings in S_{N_0}, that the values of the uth iterate along W are those of an *acceptable point* in the Fieldian construction (and that of the rank of $\mathfrak{M}$ would be the second acceptable point). Setting $\|A\| = |A|_u$ then, we get the ultimate semantic values for each $A \in L^+$,and have thus constructed the intended V^+. QED

Remark 1 Π^1_4-CA is not best possible in the last lemma. If we had assumed only Π^1_3-CA, and a kind of formal development of a constructible hierarchy over V as detailed for second order arithmetic in Simpson (1999, VII. 4), then using absolutness and relativization arguments, we could find a model of $\Pi^{1-\text{class}}_3$-CA $+ V = L[X]$, by analogy with $\Pi^{1-\text{set}}_3$-CA of Simpson (1999, VII. 3). This model would satisfy Σ^1_3-Uniformization, from which Π^1_3-REC could be shown; this, together with Lemma 5 is what is needed for the proof to go through.

References

Beall, JC (ed.) (2003). *Liars and Heaps*, Oxford University Press, Oxford

Boolos, G. (1984). 'To be is to be a value of a variable (or to be some values of some variables)' *The Journal of Philosophy* 81, 430–49. Reprinted in Boolos (1998)

——— (1985a). 'Nominalist Platonism,' *Philosophical Review* 94, 327–44. Reprinted in Boolos (1998)

——— (1985b). 'Reading the *Begriffsschrift*,' *Mind* 94, 331–4. Reprinted in Boolos (1998)

——— (1998). *Logic, Logic and Logic*, Harvard, Cambridge, Mass.

Devlin, K. (1984). *Constructibility*, Perspectives in Mathematical Logic, Springer Verlag, Berlin, Heidelberg

Field, H. (2003a). 'A Revenge-immune solution to the semantic paradoxes,' *Journal of Philosophical Logic* 32, 139–77

——— (2003b.) 'The semantic paradoxes and the paradoxes of vagueness.' In Beal (2003)

——— (2004). 'The consistency of the naive theory of properties,' *Philosophical Quarterly* 54, 78–104

Hossack, K. (2000). 'Plurals and complexes', *The British Journal for the Philosophy of Science* 51: 3 411–43

Kripke, S. (1975). 'Outline of a theory of truth,' *Journal of Philosophy* 72, 690–716

Linnebo, Ø (2003). 'Plural quantification exposed', *Noûs* 37, 71–92

McGee, V. (1997). 'How we learn mathematical language,' *Philosophical Review* 106, 35–68

_____ (2000). ' "Everything" '. In Sher and Tieszen (2000)

Moschovakis, Y. (1974). *Elementary Induction on Abstract structures*, Vol. 77 of *Studies in Logic series*, North Holland, Amsterdam

Oliver, A., and Smiley, T. (2001). 'Strategies for a logic of plurals,' *Philosophical Quarterly* 51, 289–306

Parsons, C. (1990). 'The structuralist view of mathematical objects,' *Synthese* 84, 303–46

Quine, W. V. (1986). *Philosophy of Logic*, 2nd edn., Harvard Cambridge, Mass.

Rayo, A. (2002). 'Word and objects,' *Noûs* 36, 436–64

_____ (forthcoming) 'Beyond plurals.' In Rayo and Uzquiano (forthcoming)

Rayo, A., and Uzquiano, G. (1999). 'Toward a theory of second-order consequence', *The Notre Dame Journal of Formal Logic* 40, 315–25

_____ (eds.) (forthcoming). *Absolute Generality*, Oxford University Press, Oxford

Rayo, A., and Yablo, S. (2001). 'Nominalism through de-Nominalization', *Noûs* 35: 1, 74–92

Resnik, M. (1988). 'Second-order logic still wild', *Journal of Philosophy 85: 2*, 75–87

Sher, G., and Tieszen, R. (eds.) (2000). *Between Logic and Intuition*, Cambridge University Press, New York and Cambridge

Simpson, S. (1999). 'Subsystems of second-order arithmetic', *Perspectives in Mathematical Logic*, Springer Verlag, Berlin, Heidelberg

Welch, P. (2003). 'Ultimate truth *vis à vis* stable truth'. Submitted to *Journal of Philosophical Logic*

Williamson, T. (2003). 'Everything'. 415–65. In J. Hawthorne and D. Zimmerman (eds.), *Philosophical Perspectives* 17: *Language and Philosophical Linguistics*, Blackwell, Oxford

11

Bradwardine's Revenge

Stephen Read

11.1 Bradwardine

Suppose Socrates utters sentence (A):

Socrates utters a falsehood

and I claim, using sentence (B):

Socrates utters a falsehood.

Bradwardine agrees with this assessment, but considers the following objection:

> But if it is true that Socrates utters a falsehood, and Socrates says that, then Socrates utters a truth.

His reply is:

> The minor premise is false, because Socrates does not utter the proposition proposed by you and conceded by me [that is, (B)], but another proposition just like it, namely (A).[1]

[1] [2] §7.01. (Note, however, that Roure's text in §7.01 is highly speculative and not supported by the mss. I am presently preparing a critical edition from all thirteen known mss. together with an English translation.) Thomas Bradwardine is more famous in the history of science and of theology than that of logic. However, he wrote several logical treatises when teaching in Oxford in the early 1320s. He was made Archbishop of Canterbury by the Pope in August 1349, arriving back in London on 19 August just in time to succumb to the last ravages of the Black Death in England a week later. Cf. also [16].

How can (A) be false and (B) true, when they are so similar? Again, Bradwardine considers the challenge:

> But the terms and copulae of (A) and (B) are equivalent, so (A) and (B) must be equivalent.

That does not follow, he replies. The subject and predicate of (A) and (B) refer to the same things, but (A) is self-referential while (B) is not. Hence the truth conditions of (A) and (B) are different.

Bradwardine offers us a general thesis about certain sorts of self-referential proposition. Suppose a proposition says of itself that it is false, or not true. (For Bradwardine, 'false' and 'not true' are equivalent, since he holds that every proposition is either true or false.) Then it also says of itself that it is true, and is false.[2] Why is that? It depends crucially on two principles, about *saying that* and about *truth*, which Bradwardine endorses:

> (T) A proposition is true if and only if everything it says is the case
> (K) A proposition says everything which is a consequence of what it says.

(T) is his definition of truth, and (K) is the second of six postulates (the first is Bivalence). (K) is a closure principle, that *saying that* is closed under consequence. Consider a conjunction, for example. It says what each of its conjuncts says. Or consider the case of the landowner who brings his prisoner to Sancho Panza in Cervantes' novel *Don Quixote* [3]. It is Sancho Panza's first day as governor of Barataria, and before breakfast he must preside in court. A river crosses the owner's land, and he is troubled by vagabonds and thieves, so he has decreed that anyone wishing to cross the bridge over the river must say what their business is and where they are going. Those who speak truly will be allowed to cross but those who speak falsely will die on the gallows erected beside it. All has worked well for years until a man stands before the bridge and declares:

> (C) My business is to die on the gallows and none other.

Clearly, if they let him cross the bridge he will have spoken falsely and so should be hanged, whereas if they do hang him he will have spoken the truth and so should have been allowed to cross freely.

Bradwardine's comment (some 300 years before Cervantes, for it was a familiar medieval sophism) is that 'crossing the bridge' and 'speaking the truth' are equivalent, given the facts of the case, so not speaking the truth follows from not crossing the bridge, and so from dying on the gallows. Hence, whoever says they will die on the

[2] Note that this is a weaker claim than the later claim of Buridan and others that every proposition says of itself that it is true. (See e.g. [5].) They merely asserted this. Bradwardine gave a careful and detailed proof of his more discriminating claim, which we will examine later.

gallows has implicitly said that they are not speaking the truth. Unsurprisingly then, in Bradwardine's example, it was Socrates who turned up and said 'Socrates will not cross the bridge.'

Again, it's Socrates who hears that, according to some scheme, everyone who does not utter a falsehood will receive a penny, and only them, and so he says:

(D) Socrates will not receive a penny.

Bradwardine responds:

> It must be realised that the following are equivalent according to the facts of the case: 'will not receive a penny' and 'utters a falsehood', because they are universally true of each other and because their opposites convert, namely, 'will receive a penny' and 'does not utter a falsehood'. Now (D) says that Socrates will not receive a penny, from which it follows by the rules of the scheme that Socrates utters a falsehood, and from this it follows that what was said by Socrates is a falsehood . . . and so by (K), Socrates' utterance says that it itself is false. So it is false, as before.
>
> ([2] §8.01)

Such examples should serve to make (K) plausible, and show how it works.

How does Bradwardine show that (T) and (K) entail that any proposition which says of itself that it is false also says of itself that it is true? Suppose x says of itself that it is false. It may say other things as well—given (K), it almost certainly does. Abbreviate what else it says as Q, and suppose x is false. Then by Bivalence (Bradwardine's first postulate), x is not true, so by (T), something x says fails to hold, either that x is false or Q:

$$\mathbf{False}(x) \rightarrow (Q \rightarrow \neg\mathbf{False}(x)),$$

that is, assuming x is false, and assuming that whatever else x says holds, that x is false must fail to hold, for something it says fails. So by Importation and Bivalence again,

$$(\mathbf{False}(x) \wedge Q) \rightarrow \mathbf{True}(x),$$

where '$\rightarrow$' is the residual of '$\wedge$'.[3] But x says that x is false and Q, so what x says implies that x is true. So by (K), x says that x is true. Thus we have proved the first part of Bradwardine's main thesis, that any proposition saying of itself that it is false (or not true) also says of itself that it is true.

It follows immediately that any such proposition is false, for something it says must fail to hold, either that it is false or that it is true, for we cannot have both. Thus any proposition saying of itself that it is false, such as (A), is false. So too is Socrates' assertion that he will die on the gallows, and that he will not receive a penny.

[3] As a relevance logician, I distinguish '$\wedge$' (extensional conjunction) from '$\circ$', the residual of '$\rightarrow$'. But since $P \wedge Q$ entails $P \circ Q$, it follows that $(\mathbf{False}(x) \circ Q) \rightarrow \mathbf{True}(x)$ entails $(\mathbf{False}(x) \wedge Q) \rightarrow \mathbf{True}(x)$, so the reasoning still goes through.

But it does not follow that (A) is true, or that Socrates will not die on the gallows, or that he will receive a penny. The reason (A) is false is not because Socrates spoke the truth, nor is (C) false because he would not die on the gallows, nor is (D) false because he would receive a penny. Each is false because something else which those propositions said failed to obtain, namely, that each of them is true. (C) says implicitly that it is true, since 'die on the gallows' and 'speak falsely' are equivalent, and the same with (D). So each of (C) and (D) implicitly says of itself that it is false, and so by the main thesis also says of itself that it is true. But they cannot be both true and false, so something they say cannot obtain, and so by (T) each is false.

Note that Bradwardine is not claiming that (A)'s truth follows from its being false. The above proof does not establish

$$\mathbf{False}(x) \rightarrow \mathbf{True}(x).$$

So what x says 'implicitly' does not follow from what it says 'explicitly', if one can make these distinctions. If it did, the doctrine would collapse. For Bradwardine claims that (A) is false, so if '(A) is true' followed from '(A) is false', Bradwardine would be landed in contradiction. Rather, the proof serves to tease out of what (A), or more generally x, says what it fully says. Implicit in what it says (*viz* $\mathbf{False}(x) \wedge Q$) is that it is true.

11.2 Types

It is often claimed that fourteenth-century terminists, like Ockham and Buridan, took propositions to be token utterances.[4] They certainly took them to be concrete individuals, albeit some existed only in the mind. Be that as it may, Bradwardine was an opponent of Ockham's and follower of Burley, a realist regarding universals. As true followers of Aristotle, these realists believed that such universals are manifest in their instances and multiply located, not transcendent:

> The whole universal exists in each of its particulars and is not numerically multiplied by its existence in numerically distinct particulars.[5]

So there can be many instances of (A) and (B), uttered by different individuals on different occasions. What unites these instances and makes them instances of the same type; and correlatively, what distinguishes these instances, and makes them different, so that (A) and (B) are of different types? For the same token can be an instance of different types, and so both of the same and of different type than another.

[4] See e.g. [8], *passim*; [5] pp. 5 ff.; [11] pp. 207–9.

[5] Walter Burley, *Super Artem Veterem Porphyrii et Aristotelis*, cited in [1] p. 423.

Clearly, different instances of the same grammatical type can have different truth-values, and so different truth conditions. For example,

(E) I am Greek

uttered by me and by Socrates have different truth-values, so if propositions are to be bearers of truth-values they cannot have the same identity conditions as the sentences which are used to express them. These two instances of the sentence-type (E) differ in what they say: as uttered by me it says that I am Greek, as uttered by Socrates, that he is Greek. The proposition uttered by Socrates was different from that uttered by me. His said that he was Greek, mine that I was Greek, and their truth conditions, captured by (T) as consisting in the obtaining of what each of us said, are different.

In this case, however, Socrates and I are clearly speaking about different people, himself and myself respectively. In the case of (A) and (B), that is not the case. In (A) and (B), Socrates and I are both speaking about Socrates and what Socrates said. But as Perry and Lewis showed us, that is misleading, for Socrates is speaking *de se* whereas I am not. So it is misleading to represent the content of (A) and (B) as, say,

(†)(*A*) is false,

for that overlooks the crucial self-reflexive aspect of (A) compared with (B). (A) says *de se* that it is false, whereas (B) says this *de re*. As Geach pointed out long ago ([4] §§80–4), in 'Only Satan pities Satan', and 'Only Satan pities himself', 'Satan' and 'himself' both refer to Satan, but their truth conditions are utterly different.

Perry's famous example ([13] p. 492) places Rudolf Lingens in the Main Library at Stanford, with all the world's knowledge around him, including a biography of himself and a plan of the Library: nonetheless, knowing all this, he may still fail to know that he himself is Rudolf Lingens and that he is in Stanford Library. Lewis ([10] p. 521) describes this as failure of knowledge *de se*, as contrasted with his knowledge *de dicto* of what is contained in the Library volumes. But he may have knowledge *de re* of Rudolf Lingens, as well as knowledge *de dicto*. Suppose, for example, that he is aware that someone is in the same part of the stacks: he hears their footsteps, perhaps glances at their reflection in the mirror at the end of the aisle, and so knows of someone that they are there. What he does not realize is that they are his own footsteps, echoing strangely, that it is his own reflection he sees, and so on. He has knowledge *de re* as well as *de dicto* of Lingens, but until he realizes the stranger is himself, it is not knowledge *de se*.[6]

Consequently, it does not follow from the fact that both (A) and (B) say of (A) that it is false that (A) and (B) are equivalent. We must be very careful in grouping

[6] Perry makes the same point in [14] with his example of the search for the supermarket shopper whose trolley is spilling sugar.

expressions together, and in representing what propositions say. Subtle and hidden differences, like that between (A) and (B), can easily lead to error.

We may be tempted by these reflections to propose (with Ockham, Buridan, and others—cf. n. 4) that the primary truth-bearers are tokens, and that types can only be treated as truth-bearers derivatively. But this would be a mistake, for tokens do not have the right properties to act as truth-bearers. Indeed, the situation is entirely the reverse: tokens only have properties such as truth and meaning as tokens of a type. Consider the word 'the', for example. Its tokens only have the linguistic properties they do as tokens of the type. The linguistic properties are emergent, and can only be explained as properties of the material (written or spoken) tokens by their participation in a communal practice. A similar challenge was levelled at Grice's theory of meaning by Quinton: there is a boot-strapping, whereby the linguistic properties of language, dependent though they may be on individual practice, are also required by that practice in order for the practice to succeed. I can't intend to stop the bus by tapping my kneecaps, Quinton observed ([14] p. 344). Similarly, I can't intend to stop the bus by calling out 'Stop' unless that sound has the emergent properties which it can have only as part of a linguistic practice.

The case of the word 'the' is salutary, however. Peirce, when introducing the distinction between type and token, claimed that, though there are many tokens of 'the',

> there is but one word 'the' in the English language.
>
> ([12] vol. IV §537)

Not so; the *Oxford English Dictionary* distinguishes three homophones, though one is archaic. But to make grammatical sense, we need to distinguish the demonstrative adjective 'the' from the adverb 'the', the latter instanced in such constructions as 'What student is the better for mastering these futile distinctions?', or the medieval sophism 'The uglier you are the prettier you are.' The tokens of 'the' are adjectives or adverbs respectively only in virtue of being tokens of the respective types; but the types have to take account of the linguistic function of the tokens, so that one unites in the type only those tokens with the right function.[7]

Thus (A) is self-referential and (B) is not. Note that I am speaking here of (A), not Socrates, as self-referring. So it is irrelevant whether Socrates knows he is Socrates, or referred to by the name 'Socrates'. Lingens' utterances of 'Lingens is speaking falsely' would be self-referential even though Lingens not know *de se* that he was referring to himself when he referred to Lingens. Being self-referential, and in particular, saying of itself that (A) is false, (A) falls under Bradwardine's main thesis, and so (A) also says of itself that it is true, and so is false. (B), on the other hand, is not self-referential,

[7] Faced with such phenomena, Kaplan [7] proposes abandoning the type/token distinction altogether. But what is really important are the identity criteria.

but refers to (A), which though a token of the same type-sentence, or construction, does not express the same proposition. Using Kaplan's distinction ([6]) between content and character, (A) and (B) have the same content but different character. This would be clearer had Socrates chosen to express himself by, e.g., 'I am uttering a falsehood'. But that is just, as Bradwardine would say ([2] §7.023), an 'accident'. We might represent the respective forms of (A) and (B) as:

$$(\lambda x)[x \text{ says that } x \text{ is false}](\text{A})$$
$$(\lambda x)[x \text{ says that (A) is false}](\text{B})$$

However, this description of (A) is incomplete; more fully, given Bradwardine's main thesis:

$$(\lambda x)[x \text{ says that } x \text{ is false} \wedge x \text{ says that } x \text{ is true}](\text{A})$$

So unlike (A), (B) does not say of itself that it is false, and so need not be false. Indeed, by Bradwardine's result, it is true. (A) is indeed false.

11.3 Truth

Bradwardine's main thesis applies to any proposition which says of itself that it is false. So it provides a solution not just to the traditional Liar paradox, as exemplified in (A), but to many other related paradoxes. Indeed, the medievals were especially fascinated by the range of logical paradoxes, or 'insolubles' as they were known. Among 'insolubles' which Bradwardine treats are the conjunctive and disjunctive cases:

(G) God exists and a false conjunction is uttered by Socrates

and

(H) A man is an ass or a false disjunction is uttered by Socrates

(each taken to be the only conjunction or disjunction uttered by Socrates, and where 'God exists' is taken to be true and 'A man is an ass' false).[8] Take (G), for example, and call its second conjunct (I). Then (I) says that (G) is false, since (G) is assumed to be the only conjunction Socrates utters. But if (G) is false, then one of its conjuncts is false, so it follows from what (I) says (namely, that (G) is false) that (I) is false. So by (K), (I) says of itself that it is false, and so, by the main thesis, says of itself that it is true, and consequently is false. Hence (G) is false, since it has a false conjunct. A similar analysis applies to (H).

From here, it is straightforward to extend the analysis to, e.g., Curry's paradox:

(J) If (J) is true then a man is an ass

[8] [2] §§8.05–06.

(for (J) is equivalent to a disjunction similar to (H)), and to further paradoxes such as Yablo's paradox, the postcard paradox (in two forms), and the truth-teller puzzle, as I have done elsewhere ([19]). What is crucial to realize is that adopting the combination of (T) and (K) disables the second leg (see below) of the standard proofs of a contradiction from admitting the construction of the so-called paradoxes. Take the Liar paradox, in its form (A) above. One starts by assuming (A) is true, from which it follows by (T) that (A) is false, so by *reductio* (A) is false. So far so good—that's the first leg. But (this is the second leg) if (A) is false, then surely, since that is what it says, (A) must be true, yielding the familiar contradiction. But (T) does not justify this reasoning. According to (T), it is not sufficient for the truth of (A) that (A) be false. The truth of (A) requires that everything that (A) says obtain, and by (K), (A) says not only that (A) is false but also, we saw in Bradwardine's main thesis, that (A) is true. So the proof that (A) is true fails, and the reason it fails has been diagnosed; but more: the reason it was thought to succeed has also been revealed. For (A) appears to say only that (A) is false. That it also, implicitly, says that (A) is true is hidden and far from obvious. So one can easily be misled into thinking that the reasoning concluding that (A) is both true and false is valid. Bradwardine's solution not only diagnoses what is wrong, but also explains why the mistake was made.

Elsewhere, ([17] and [18]) I have described (T) as a revision or challenge to Tarski's famous T-scheme. Although I stand by that charge, it is worth looking a little more closely at Tarski's reasoning concerning the T-scheme. To begin with, we should note that whereas (T) is intended as a definition of truth, Tarski explicitly denied that the T-scheme was such a definition ([22] p. 357). Rather, the T-scheme was presented as a material adequacy condition, that is, a test of the extensional correctness of a putative definition. The definition of truth which Tarski gave consisted in a recursive definition of satisfaction culminating in the definition:

A sentence is true if it is satisfied by all objects, and false otherwise.[9]

But more interesting is the reasoning through which Tarski proceeded to his test of material adequacy.

Tarski starts from the famous Aristotelian formula in *Metaphysics* Γ 7,[10] which he expresses in more contemporary idiom as

(*) The truth of a sentence consists in its agreement with (or correspondence to) reality.

([22] p. 343)

But he dismisses this as unsatisfactory for two reasons. First, it is not 'sufficiently precise and clear'; secondly, it leads to paradox. What is needed, he says, is to take (*)

[9] [22] p. 353; cf. Definition 23 [21] p. 195.

[10] 'To say of what is that it is not or of what is not that it is, is false, while to say of what is that it is, or of what is not that it is not. is true.' (1011b25)

and make it 'more definite, and . . . give it a correct form' ([21] p. 155). He starts with what he calls 'partial definitions', defining truth for each sentence *seriatim*. The general form of such a definition is

(P) *x* is a true sentence if and only if *p*,

where 'we substitute in the place of the symbol "*p*" in this scheme any sentence, and in place of '*x*' any individual name of this sentence' (pp. 155–6).

First, note Tarski's vehement rejection of the suggestion of a 'defect' in this scheme, the claim that it needs supplementation: ' . . . if and only if *p* is true' or ' . . . if and only if *p* is the case (i.e. if what *p* states is the case)' ([22] p. 357). This is 'based upon an obvious confusion between sentences and their names' (p. 358). What replaces '*p*' in (P) is a sentence, so appending 'is true' or 'is the case' to it is nonsense. (P) is a schema, of which

'Snow is white' is true if and only if snow is white

is the most famous instance. 'if and only if' is a sentential connective which connects sentences—it does not connect names of sentences. '*x*' is a variable over names and '*p*' is a variable over sentences—that is, what replaces '*x*' is a name and what replaces '*p*' is a sentence (and *not* a name).

Note, too, that '*p*' is a sentence-variable, not a dummy letter. Tarski does not hesitate to quantify over sentence-variables. To be sure, he rejects

$$(\forall x)(x \text{ is a true sentence if and only if } (\exists p)(x = \ulcorner p \urcorner \text{and } p)),$$

but not on the grounds that it is not well formed (at least, not once $\ulcorner p \urcorner$ is correctly interpreted as a quotation name containing '*p*' as a free variable, not as a name of the letter '*p*'),[11] but as opening the door to paradox. After all, here was a student of Leśniewski's whose doctoral dissertation (summarized in [20]) contained as its Fundamental Theorem:

$$(\forall p, q)(p \wedge q \equiv (\forall f)(p \equiv ((\forall r)(p \equiv f(r)) \equiv (\forall r)(q \equiv f(r))))),$$[12]

expressed in the type theory of *Principia Mathematica*. The objection to (P) was not that one could not generalize the partial definitions, but that that generalization led to paradox. The paradox was illustrated by, in effect, replacing '*p*' by '(A) is false' and '*x*' by '(A)', to obtain

(‡) (A) is true if and only if (A) is false.

This contrasts intriguingly with the way Tarski proceeds to incorporate (P) into his notorious 'Convention **T**'. Where in (P) Tarski requires (as cited above) that we

[11] [21] p. 161. [12] [20] p. 2.

replace '*x*' by any name of the sentence which replaces '*p*', in Convention **T** Tarski proposes as his test of extensional correctness:

Convention **T**. *A formally correct definition of the symbol 'Tr', formulated in the metalanguage, will be called an* adequate definition of truth *if it has the following consequences*:

(α) *all sentences which are obtained from the expression 'x* $\in$ *Tr if and only if p' by substituting for the symbol 'x' a structural-descriptive name of any sentence of the language in question and for the symbol 'p' the expression which forms the translation of this sentence into the metalanguage*;

(β) ...

([21] pp. 187–8)

It is easy to read this as simply adapting (P) from the homophonic (and so unacceptable) case of a semantically closed language where the sentence whose truth is in question replaces '*p*' to the new situation where truth is defined in the metalanguage, and what replaces '*p*' is the metalinguistic correlate of the object-language sentence in question. But one can read it more generously. For example, take a sentence containing an indexical, like (E), and apply Convention **T**. Then if the metalanguage is mine, but the utterance of (E) is Socrates', I should clearly require that my truth theory entail

'I am Greek' is true if and only if Socrates is Greek,

and not

'I am Greek' is true if and only if I am Greek,

for in the latter I have not properly translated Socrates' utterance into my metalanguage.

If Bradwardine is right about what (A) says, then again, correctly to apply Convention **T**, I have to translate (A) properly into my metalanguage. But (‡) does not do that, for it omits the crucial consequence of the self-referential nature of (A), namely, that it says not only that (A) is false but also that (A) is true. So the proper instance of Convention **T** is

(A) is true if and only if, (A) is false and (A) is true.

But that will only be a consequence of a truth theory which makes (A) false (or at least, not true).

Understood in this way, Tarski's Convention **T** does give the correct material adequacy condition, but then does not support Tarski's rejection of semantic closure as paradoxical. What replaces '*p*' must express everything which the sentence in question, whose name replaces 'x', says, as is required in (T). Hence Bradwardine's theory of truth, (T), meets Tarski's adequacy condition, properly understood. That theory shows that (A) is false, and (B) is true, and so is immune to revenge.

11.4 Conclusion

Bradwardine claimed that any proposition which says of itself that it is false, also says of itself that it is true, and so, since things cannot be wholly as it says they are (namely, both true and false), it is false. He proved this, his main thesis, from assumptions foremost among which was the closure principle, that every proposition says whatever follows from what it says.

This thesis applies to a whole range of apparent paradoxes, among them the classic case of the Liar paradox. Bradwardine goes on to consider the notorious problem of 'revenge': if one wishes to claim that the Liar sentence is false, has one not thereby committed oneself to claiming that it is true that it is false, and so shown that it is true? Bradwardine's defence hinges on distinguishing the claim that the Liar is false from the Liar itself. These propositions may be superficially indistinguishable, for they may be expressed by the same type-sentence, indeed, they may seem semantically identical, in both saying of the Liar that it is false. But the Liar says this of itself, thereby falling under Bradwardine's main thesis and so saying of itself both that it is false and that it is true. The theorist's claim that the Liar is false is not self-referential in this way, and so is simply false.

Tarski claimed that such self-reference leads inevitably to paradox, by substituting the Liar into his material adequacy condition. But now that we realize that the Liar says more than just that it is false we see that that substitution was not carried through properly. The T-condition requires that the truth of any sentence equate to a statement of what it says. The condition was not presented by Tarski as a theory of truth but as an extensional test of the adequacy of any such theory. We now see that Bradwardine's theory of truth meets that test: for the Liar is true if and only if what it says is the case, namely, if it is both true and false. It cannot be both, so it is false. But the statement that it is false is true, for that statement is not itself self-referential, but says no more than that the Liar is false and whatever follows from that.

References

[1] Adams, M. A. (1982). 'Universals in the early fourteenth century'. In Kretzmann *et al.* (1982) 411–39

[2] Bradwardine, T. (1970). *Insolubilia*, in M.-L. Roure (1970). 'La problématique des propositions insolubles au XIII^{e} siècle et au début du XIV^{e}, suivie de l'édition des traités de W. Shyreswood, W. Burleigh et Th. Bradwardine', *Archives d'Histoire Doctrinale et Littéraire du Moyen Age* 37, 205–326

[3] Cervantes, M. de (1986). *Don Quixote*, tr. T. Smollett. London: Deutsch

[4] Geach, P. T. (1962). *Reference and Generality*. Ithaca: Cornell University Press
[5] Hughes, G. E. (1982). *John Buridan on Self-reference. Chapter Eight of Buridan's 'Sophismata', translated with an Introduction, and a philosophical commentary*. Cambridge: Cambridge University Press
[6] Kaplan, D. (1979). 'On the logic of demonstratives', *Journal of Philosophical Logic* 8, 81–98
[7] ——— (1990). 'Words', *Aristotelian Society* Supplementary vol. 64, 93–119
[8] Klima, G. (2004). 'Consequences of a closed, token-based semantics: the case of John Buridan', *History and Philosophy of Logic* 25, 95–110
[9] Kretzmann, N., Kenny, A., and Pinborg, J. (eds.) (1982). *The Cambridge History of Later Medieval Philosophy*. Cambridge: Cambridge University Press
[10] Lewis, D. K. (1979). 'Attitudes *de dicto* and *de se*', *The Philosophical Review* 88, 513–43
[11] Nuchelmans, G. (1982). 'The semantics of propositions'. In Kretzmann *et al.* (1982) pp. 197–210
[12] Peirce, C. S. (1932). *Collected Papers*, ed. C. Hartshorne, P. Weiss, and A. W. Burks Cambridge, Mass.: Harvard University Press
[13] Perry, J. (1977). 'Frege on demonstratives', *The Philosophical Review* 86, 474–97
[14] ——— (1979). 'The problem of the essential indexical', *Nous* 13, 3–21
[15] Quinton, A. M. (1973). *The Nature of Things*. London : Routledge Kegan Paul
[16] Read, S. (2002). 'The Liar paradox from John Buridan back to Thomas Bradwardine', *Vivarium* 40, 189–218.
[17] ——— (2006). 'The Truth schema and the Liar'. Forthcoming in S. Rahman and T. Tulenheimo (eds.), *Unity, Truth and the Liar*. Berlin: Springer
[18] ——— (2006a). 'Further thoughts on the T-scheme and the Liar'. Forthcoming in S. Rahman and T. Tulenheimo (eds.), *Unity, Truth and the Liar*. Berlin: Springer
[19] ——— (2006b). 'Symmetry and paradox', *History and Philosophy of Logic*, forthcoming
[20] Tarski, A. (1927). 'On the primitive term of logistic', in Tarski (1956) 1–23
[21] ——— (1936). 'The concept of truth in formalized languages'. In Tarski (1956) 152–278
[22] ——— (1944). 'The semantic conception of truth', *Philosophy and Phenomenological Research* 4, 341–76
[23] ——— (1956). *Logic, Semantics, Metamathematics*, trans. J. Woodger. Oxford: Clarendon

12

Curry's Revenge: The Costs of Non-classical Solutions to the Paradoxes of Self-reference

Greg Restall

The paradoxes of self-reference are *genuinely* paradoxical. The liar paradox, Russell's paradox, and their cousins pose enormous difficulties to anyone who seeks to give a comprehensive theory of semantics, or of sets, or of any other domain which allows a modicum of self-reference and a modest number of logical principles.

One approach to the paradoxes of self-reference takes these paradoxes as motivating a *non-classical* theory of logical consequence. Similar logical principles are used in each of the paradoxical inferences. If one or other of these problematic inferences are rejected, we may arrive at a consistent (or at least, a coherent) theory.

In this chapter I will show that such approaches come at a serious cost. The general approach of using the paradoxes to restrict the class of allowable inferences places severe constraints on the domain of possible propositional logics, *and* on the kind of metatheory that is appropriate in the study of logic itself. Proof-theoretic and model-theoretic analyses of logical consequence make provide different ways for non-classical responses to the paradoxes to be defeated by *revenge* problems: the redefinition of logical connectives thought to be ruled out on logical grounds. Non-classical solutions are not the 'easy way out' of the paradoxes.[1]

[1] It does not follow that non-classical accounts of the paradoxes are misguided or wrong-headed. On the contrary, I think that the general approach is quite sane, and have argued as much in print [18].

12.1 Non-Classical Solutions

In this section I will sketch the structure of non-classical approaches to the paradoxes. They have straightforward general features: Firstly, we keep whatever semantic, or set-theoretic principles are at issue. For example, if it is the liar paradox in question, we can keep the naïve truth scheme, to the effect that

$$T\langle A\rangle \leftrightarrow A$$

where $\langle\text{—}\rangle$ is some name-forming functor, taking sentences to names, and where $\leftrightarrow$ is some form of biconditional. This scheme says, in effect, that $T\langle A\rangle$ is true under the same circumstances as A. To assert that A is true is saying no more and no less than asserting A.

Secondly, we allow our language to contain a modicum of self-reference. We wish to express sentences such as the liar: 'This very sentence is not true.' If the language in question is a natural language, then indexicals will do the trick. If the language is a formal language without indexicals, some other technique will be needed to construct sentences analogous to the liar. A Gödel numbering and a means of diagonalization will do nicely to give the required results.[2]

With such machinery at hand, we can reason as follows: Use a means of diagonalization to construct a statement λ such that λ is equivalent to $\sim T\langle\lambda\rangle$. Then reason as follows: $\lambda \leftrightarrow \sim T\langle\lambda\rangle$, but by the T-scheme, $\lambda \leftrightarrow T\langle\lambda\rangle$. Therefore $T\langle\lambda\rangle \leftrightarrow \sim T\langle\lambda\rangle$, and equivalently, $\lambda \leftrightarrow \sim\lambda$. We can then deduce $\lambda \wedge \sim\lambda$ (from an inference such as *reductio*: $p \rightarrow \sim p \vdash \sim p$) and we have a contradiction.

If your favourite paradox is Russell's, instead of the liar, the non-classical approach will keep the naïve class abstraction scheme

$$x \in \{y : \phi(y)\} \leftrightarrow \phi(x)$$

and you reason similarly, from the definition of the Russell class r as $\{x : x \notin x\}$. If $r \in r$ then $r \notin r$, and if $r \notin r$ then $r \in r$. The same holds for Berry's paradox, the Burali–Forti paradox, and many others.[3]

The non-classical response to these paradoxes is to find fault with the logical principles involved in the deduction. Most approaches to the paradoxes take them to be important lessons in the behaviour of *negation*. There are two different lessons we might learn. One is that the inference from $A \leftrightarrow \sim A$ to $A \wedge \sim A$ fails, since A might be (speaking crudely) neither true nor false. Another possible lesson is that the inference from $A \wedge \sim A$ to an arbitrary B fails, since A can be (speaking less crudely

[2] See Boolos and Jeffrey [4] for a review of the standard approach, and see Smullyan [21] for more on what a language must contain to feature self-reference.

[3] A compendium of such paradoxes is given by Graham Priest [17].

this time) both true and false. However negation works, it cannot be *Boolean*. Boolean negation allows *both* inferences, and inferring *every* statement from a the existence of liar sentence or a Russell set is just too much. Boolean negation is rather too strong, so an alternative logic of negation must be found [5, 6, 16, 17].

If you wish to define negation non-classically, there are many options available. You can define negation *inferentially*, taking $\sim A$ to mean that *if* A, *then* something absurd follows,[4] or it can be defined by way of the equivalence between the *truth* of $\sim A$ and the *falsity* of A, and allowing truth and falsity to have rather more independence from one another than is usually taken to be the case: say, allowing statements to be *neither* true nor false, or *both* true and false.[5] The former account takes truth as primary, and defines negation in terms of a rejected proposition and implication. In the context of the semantics of relevant logics, this approach is sometimes called the *Australian Plan*. The account which takes truth and falsity as on a par is sometimes called the *American Plan*. In either case, there are many options for the theorist seeking an alternative account of negation.

I sketch this general typology of negation merely to indicate that I need not take a stand on it. In what follows we will see that the paradoxes have more to teach us than this. If we wish to be non-classical, we need to work with much more than the logic of negation.

12.2 Curry's Paradox

The paradox I have in mind can be found in a logic independently of its stand on negation. The deduction appeals to no particular principles of negation, as it is negation-free. Any deduction must use some inferential principles. Here are the principles needed to derive the paradox.

A Transitive Relation of Consequence: We write this by '$\vdash$'. I take $\vdash$ to be a relation between statements, and I require that it be transitive: if $A \vdash B$ and $B \vdash C$ then $A \vdash C$.

Conjunction and implication: I require that the conjunction operator $\wedge$ be a greatest lower bound with respect to $\vdash$. That is, $A \vdash B$ and $A \vdash C$ if and only if $A \vdash B \wedge C$. Furthermore, I require that there be a *residual* for conjunction: a connective $\supset$ such that

$$A \wedge B \vdash C \text{ if and only if } A \vdash B \supset C$$

[4] See, e.g. Meyer and Martin's account of negation as implying falsehood, and its idiosyncrasies when combined with a *relevant* notion of implication [11].

[5] Three examples are four valued semantics of relevant logics, used by Dunn [8, 9] and Belnap [17], the semantics of Priest's *In Contradiction* [15], and the semantics of Nelson's constructible falsity [15] and its extensions by Heinrich Wansing [23].

This is our connective of implication. (You may wonder how we might come across such a connective. There are many ways to construct it. In the next section we will examine some.)

A paradox generator: We need only a very weak paradox generator. We take the T scheme in the following *enthymematic* form:

$$T\langle A\rangle \wedge C \vdash A \qquad A \wedge C \vdash T\langle A\rangle$$

for some true statement C. The idea is simple: $T\langle A\rangle$ need not *entail* A.[6] Take C to be the conjunction of all required background constraints. It is true that the *sentence* 'snow is white' could be true without snow being white. However, if 'snow is white' is true and some background semantic theory is holds, then it follows that snow is white. Conversely, if snow is white and the background semantic theory holds, then 'snow is white' is true. Let C be that background semantic theory. It must simply give A under some background constraints (such as some facts about language) we can infer A. (If you find it difficult to construct the required background semantic theory, do not worry. Take C to be the conjunction of *all* truths: a maximally specific statement. Then we need simply that there is no instance of A for which in which $T\langle A\rangle$ is true and A fails to be true, or vice versa.)

Diagonalization: To generate the paradox we use a technique of diagonalization to construct a statement λ such that λ is equivalent to $T\langle\lambda\rangle \supset A$, where A is any statement you please. Then, with this A chosen, we reason as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{C \wedge T\langle\lambda\rangle \vdash \lambda \quad \lambda \vdash T\langle\lambda\rangle \supset A}{C \wedge T\langle\lambda\rangle \vdash T\langle\lambda\rangle \supset A}}{C \wedge T\langle\lambda\rangle \wedge T\langle\lambda\rangle \vdash A}(*)}{C \wedge T\langle\lambda\rangle \vdash A}}{C \vdash T\langle\lambda\rangle \supset A} \qquad \dfrac{T\langle\lambda\rangle \supset A \vdash \lambda}{C \vdash \langle\lambda\rangle} \qquad \dfrac{C \wedge \lambda \vdash T\langle\lambda\rangle}{C \vdash T\langle\lambda\rangle} \qquad \dfrac{\text{from } (*)}{C \wedge T\langle\lambda\rangle \vdash A} \qquad \dfrac{}{T\langle\lambda\rangle \vdash C \supset A}}{\dfrac{C \vdash C \supset A}{\dfrac{C \wedge C \vdash A}{C \vdash A}}}$$

This is a problem. Our true C entails an arbitrary A.

This inference arises independently of any treatment of negation. The form of the inference is reasonably well known. It is *Curry's paradox*, and it causes a great

[6] So, we need not take the range of the T-scheme to be *propositions* [10], as we do not need to commit ourselves to the *equivalence* of $T\langle A\rangle$ and A.

deal of trouble to any non-classical approach to the paradoxes [12, 13, 14, 20]. In the next section I show how the tools for Curry's paradox are closer to hand than you might think. Avoiding this paradox severely constrains the non-classical theorist.

12.3 The Revenge Problem

There are many different ways to get the logical tools necessary for our problematic deduction. In particular, there are many ways to get a connective $\supset$ which residuates conjunction. We will examine them one at a time.

Boolean negation: If Boolean negation is present (write it '$\sim$') then we can define $A \supset B$ to be $\sim A \vee B$. However, the non-classical theorist has explicitly rejected Boolean negation, so we need not tarry here. This is not a problem by itself.

Intuitionistic logic: The rule for the residual is satisfied by the conditional of intuitionistic logic. Any semantic account which motivates intuitionism motivates the residual of conjunction. Now no non-classical theorist of the paradoxes is going to *explicitly* use the intuitionistic conditional, for it is well known to suffer from Curry-style paradoxes. Our point in the rest of the chapter is to show that the *implicit* acceptance of this conditional is deeply embedded in our practices of logic.

Infinitary disjunction: A Curry-paradoxical conditional can arise as a revenge problem for the non-classical theorist without *explicitly* motivating intuitionistic implication. If we have *infinitary* disjunction at hand, such that a (finite) conjunction distributes over infinitary disjunction, we can define $B \supset C$ to be

$$\bigvee\{A : A \wedge B \vdash C\}$$

This will satisfy the definition of $\supset$. If $A' \wedge B \vdash C$ then $A' \vdash \bigvee\{A : A \wedge B \vdash C\}$, since $A' \in \{A : A \wedge B \vdash C\}$. Conversely, if $A' \vdash B \supset C$, we have $A' \vdash \bigvee\{A : A \wedge B \vdash C\}$. Then $A' \wedge B \vdash B \wedge \bigvee\{A : A \wedge B \vdash C\}$ and by the distribution of conjunction over disjunction, $A' \wedge B \vdash \bigvee\{A \wedge B : A \wedge B \vdash C\}$ and clearly $\bigvee\{A \wedge B : A \wedge B \vdash C\} \vdash C$, so $A' \wedge B \vdash C$ by the transitivity of entailment. Therefore, *any* semantic theory which motivates infinitary disjunction and distributive lattice logic motivates the residual $\supset$ of conjunction, and our problematic inference. This seriously constrains non-classical solutions to the paradoxes, for infinitary disjunction can be motivated in many different ways.

Proof Theory: If your favoured way to introduce connectives is by way of natural deduction (introduction and elimination rules) then infinitary disjunction is no less

motivated than ordinary disjunction. To infer $\bigvee X$ from a statement A, it is sufficient to infer a member of X.

$$\frac{A \vdash B_i}{A \vdash \bigvee\{B_i : i \in I\}}$$

If you can infer A from each element of X, then you can infer A From $\bigvee X$ too.

$$\frac{A_i \vdash B \text{ (each } i \in I)}{\bigvee\{A_i : i \in I\} \vdash B}$$

This rule is the left-hand Gentzen rule. For a traditional elimination rule for a natural deduction system, you use

$$\frac{C \vdash \bigvee\{A_i : i \in I\} \quad A_i \vdash B \text{ (each } i \in I)}{C \vdash B}$$

which is equivalent, given the transitivity of entailment. These rules seem to motivate the connective straightforwardly. However, a non-classical theorist of the paradoxes must do one of two things. One response is to allow the connective but to deny the distribution of conjunction over disjunction: that is, we do not have

$$A \wedge \bigvee\{B_i : i \in I\} \vdash \bigvee\{A \wedge B_i : i \in I\}$$

Such an approach has its own difficulties: however, it may be attempted. The second response is to reject the definition of $\bigvee$ in some way. It must be argued that this does not define a connective. This will require giving a precise account of what proof-theoretical principles are *permissible* in the account of a logical connective, and which principles are illicit. To avoid doing this is to leave the theory open to Curry's revenge.

The problem does not end here, however. The non-classical theorist must also have something to say in areas other than proof theory, for we can define disjunction in many different ways.

The Algebra of Propositions: Some logicians treat the class of propositions as an *algebra*. This algebra is closed under various operations, which have different algebraic properties. The algebra of propositions is *complete* if it is closed under arbitrary conjunctions and disjunctions. The non-classical theorist (who accepts the distribution of conjunction over disjunction) must hold that the 'intended' algebra of propositions is incomplete. This is not a particularly great burden in and of itself. However, it becomes a burden when we consider the constructions available which naturally *complete* incomplete lattices [7, 19].

Here is one result of this nature. If the lattice of propositions is incomplete, then define a *new* lattice of propositions like this. The new propositions are *ideals* of the old lattice. A set I of propositions is an ideal if and only if it is closed under converse entailment (if a entails b and $b \in I$ then $a \in I$ too) and disjunction (if $a \in I$ and $b \in I$

then $a \vee b \in I$ too). You can think of I as a set of propositions such that you would like *one* of them to be true. Our conditions ensure that we add into the set any other proposition such that making it true will be enough to make one of our original choices true. (The smallest ideal containing $\{a, b, c\}$ is the set of all propositions entailing $a \vee b \vee c$. If any of these propositions are true, then $a \vee b \vee c$ is true, which ensures that either a or b or c is true.)

Now the ideals behave *just like propositions*. The conjunction of a class of ideals is the intersection of that class. The disjunction of a class of ideals is the smallest ideal containing that class. The entailment relation among ideals is just the relation of inclusion. The collection of ideals forms a *complete* lattice. Every set of ideals has a disjunction and a conjunction. The logic of the set of ideals is very similar to the logic of propositions out of which it was constructed. However, it is complete.[7]

This is not merely a mathematical construction with no possibility for interpretation. Given an algebra of propositions, any ideal in the structure can be treated as a proposition, with simple truth conditions: I is *true* just when one member of I is true. Given a class I of propositions, it makes sense to commit yourself to the claim that one member of I is true, and this claim ought to be true just when one member of I is true.[8]

To avoid revenge, the non-classical theorist explain the point at which this reasoning breaks down. There is some ideal in the structure such that the truth of a member of I is not expressible in the domain of propositions. This is a strange result indeed, and it is a cost to the non-classical theory. It seems that the class of propositions of the language must forever remain incomplete.[9]

Once this version of the revenge problem is avoided, there yet is *another* way in which Curry's paradox might take its revenge.

State Models: Perhaps the simplest way to construct infinitary disjunction is by way of what we might call 'state models' of our logics. In a state model, each proposition is modelled by the set of states in which that proposition is true. Possible worlds models for modal logics are one form of state model.

Given a state model, it seems that infinitary disjunction is close at hand. Take a class of propositions. Their disjunction is true at the union of the class of sets of states at which each proposition in that class is true. The disjunction is true at a state just when

[7] The proof is not difficult, but I will not rehearse it here [7, 19].

[8] It would be very odd for a non-classical logician to reject *this* step, for she is the one defending the equivalence of $T\langle A\rangle$ and A.

[9] It will not do, either, to say that there are too many ideals to be expressible in a finitary language. For we do not *need* infinitary expressions to justify the existence of the residual $\supset$: The only infinitary disjunctions we need are those of the form $\bigvee\{A : A \wedge B \vdash C\}$ and these can be expressed in a finitary fashion.

one member of the disjunction is true at that state. This will define a proposition, which is the infinitary disjunction in the language. This construction relies on the notion that this class of states gives rise to a proposition. The non-classical theorist is free to reject this. However, to do so would require an explanation of which classes of states *do* give rise to propositions and which do not, and to explain why it rules out the kind of infinitary disjunction sufficient for generating the conditional for Curry's revenge.

12.4 Choices

Here, then, are the choices for any theory which seeks to give an account of the paradoxes of self-reference.

Reject large disjunctions: This requires formulating responses to each of the arguments of the previous section. This has not been done, as yet, and it is unclear what a non-classical theory which takes those arguments seriously might look like. In particular, it would aid the cause of the non-classical theorist to be able to point to a particular class of propositions and to explain why *that* class has no disjunction. It is unclear what such an explanation could look like.

Reject distribution: A crucial step in each argument has been the distribution of conjunction over disjunction. This inference has been under question for a number of reasons; primarily in quantum logic and in substructural logics. It is unclear how to motivate the failure of distribution in *this* context. It would be very nice to be able to point to a particular case of distribution and to have an explanation of why the premise is true but the conclusion fails. Such explanations are forthcoming in quantum logic (even if they are not always convincing). We need one to motivate the failure of distribution for non-quantum reasons. Uwe Petersen has the most fully developed non-classical theory of the paradoxes which lives without distribution [16]. However, no-one has given an explanation of *why* distribution fails, other than as an artefact of the proof theory. J. L. Bell has developed a semantics for quantum logic which motivates the failure of distribution [1]. It would be a great advance too if such a semantics could help *explain* a failure of distribution in the case of the paradoxes.

Reject the transitivity of entailment: This may be seen to be cutting off one's nose to spite one's face, but this approach has its proponents. Neil Tennant gives one theory of consequence which abandons the transitivity of entailment [22]. Tennant does not do this for reasons of the paradoxes, and Tennant's non-transitive logical systems do not seem to help in this case. For Tennant, if we have a proof from A to B and a proof from B to C is valid, then we do not necessarily have a proof from A to

C, but we *do* have either a refutation of *A* (a proof from *A* to the empty conclusion) or a proof of *C* (a proof of *C* from the empty premises) or a proof from *A* to *C*. So, for our purposes we may talk of arguments being *weakly* valid if and only if there is a Tennant-proof of some *superset* of the premises to some *superset* of the conclusions, and this notion of consequence is transitive, and it does all that we need to generate the paradoxes. Avoiding the transitivity of consequence in Tennant's style does not suffice for avoiding Curry's revenge.

Reject the strong laws: To live without the *T*-scheme or the naïve class comprehension scheme is to give up the goal of giving a non-classical account of the paradoxes. If the fault *isn't* with the logic but is with the semantics or the mathematics or whatever else we used, then the paradoxes do not motivate a non-classical logical theory. A classical one will do.

Each approach has its cost. None is straightforward. There is much work left to do, if we wish to give *any* account of the account of the paradoxes, including those which involve revising logical theory.[10]

References

[1] Bell, J. L. (1986). 'A new approach to quantum logic'. *British Journal for the Philosophy of Science* 37: 83–99

[2] Belnap, Nuel D. (1977). 'How a computer should think'. In G. Ryle (ed.), *Contemporary Aspects of Philosophy*. Oriel Press

[3] ——— 'A useful four-valued logic'. In J. Michael Dunn and George Epstein (eds.), *Modern Uses of Multiple-valued Logics*, pp. 8–37. Reidel, Dordrecht

[4] Boolos, George and Jeffrey, Richard (1989). *Computabilit and Logic*. Oxford University Press, 3rd edn.

[5] Brady, Ross T. (1983). 'The simple consistency of a set theory based on the logic csq'. *Notre Dame Journal of Formal Logic* 24: 431–49

[6] Brady, Ross T., and Routley, Richard (1989). 'The non-triviality of extensional dialectical set theory'. In Graham Priest, Richard Routley, and Jean Norman (eds.), *Paraconsistent Logic: Essays on the Inconsistent*, pp. 415–36. Philosophia Verlag

[7] Davey, B. A., and Priestle, H. A. (1990). *Introduction to Lattices and Order*. Cambridge University Press, Cambridge

[8] Dunn, J. Michael (1971). 'An intuitive semantics for first degree relevant implications' (abstract). *Journal of Symbolic Logic* 36: 363

[10] Thanks to audiences at the 1998 Australasian Association of Philosophy Conference and at Stanford University (especially John Etchemendy, Grigori Mints, Chris Mortensen, and Graham Priest) for helpful comments on earlier versions of this chapter.

[9] ____ (1976). 'A Kripke-style semantics for *R*-mingle using a binary accessibility relation'. *Studia Logica* 35: 163–72

[10] Horwich, Paul (1990). *Truth*. Basil Blackwell, Oxford

[11] Meyer, Robert K., and Martin, Errol P. (1986). 'Logic on the Australian Plan'. *Journal of Philosophical Logic* 15: 305–32

[12] Meyer, Robert K., Routley, Richard, and Dunn, J. Michael (1979). 'Curry's paradox'. *Analysis* 39: 124–8

[13] Myhill, J. (1975). 'Levels of implication'. In A. R. Anderson, R. C. Barcan-Marcus, and R. M. Martin (eds.), *The Logical Enterprise*, pp. 179–85. Yale University Press, New Haven

[14] ____ (1984). 'Paradoxes'. *Synthese*. 60: 129–43

[15] Nelson, D. (1949). 'Constructible falsity'. *Journal of Symbolic Logic* 14: 16–26

[16] Petersen, Uwe (1992). 'Diagonal method and dialectical logic' (unpublished book)

[17] Priest, Graham (1987). *In Contradiction: A Study of the Transconsistent*. Martinus Nijhoff, The Hague

[18] Restall, Greg (1993). 'Deviant logic and the paradoxes of self reference'. *Philosophical Studies* 70: 279–303

[19] ____ (2000). *An Introduction to Substructural Logics*. Routledge

[20] Shaw-Kwei, Moh (1954). 'Logical paradoxes for many-valued systems'. *Journal of Symbolic Logic* 19: 37–40

[21] Smullyan, Raymond M. (1994). *Diagonalization and Self-reference*, Vol. 27 of *Oxford Logic Guides*. Clarendon Press

[22] Tennant, Neil (1994). 'The transmission of truth and the transitivity of deduction'. In Dov Gabbay (ed.), *What is a Logical System?*, Vol. 4 of *Studies in Logic and Computation*, pp. 161–77 Oxford University Press, Oxford

[23] Wansing, Heinrich (1993). *The Logic of Information Structures*. No. 681 in Lecture Notes in Artificial Intelligence. Springer-Verlag

13

Aletheic Vengeance[1]

Kevin Scharp

Before you set out for revenge, first dig two graves.
attributed to Confucius

13.1 Introduction

Thinking about truth can be more dangerous than it looks. Of course, our concept of truth is the source of one of the most frustrating and impenetrable paradoxes humans have ever contemplated, the liar paradox, but that is just the beginning of its treachery. In an effort to understand why one of the most beloved and revered members of our conceptual repertoire could cause us so much trouble, philosophers have for centuries proposed 'solutions' to the liar paradox. However, it seems that our concept of truth takes offense to our efforts to understand it because it appears to retaliate against those who propose 'solutions' to the liar. It takes its revenge on us by creating new paradoxes from our own attempts to find resolution. That is, most proposed solutions to the liar paradox give rise to new, more insidious paradoxes—often called *revenge paradoxes*. For our attempts at understanding, truth rewards us with inconsistent theories, untenable logics, and a deep feeling of bewilderment. It is as if our concept of truth lashes out at us because it wants to remain a mystery. After a few run-ins with truth, many philosophers have the good sense to keep their distance. Far from being

[1] I use 'aletheic' as an adjective meaning *pertaining to truth*.

the serene, profound concept most people take it to be, those of us who think much about the liar paradox know truth to be a vengeful bully—a conceptual misanthrope.

Why has truth treated us in this way? And is there anything we can do about its misdeeds? I suggest that part of the blame falls on us, for there is a reason it is angry with us; truth is a bully, but it isn't a sociopath. Its wrath is partly a result of our insensitivity. We have tried to impose our conceptual will on truth; we have been unwilling to accept it for what it is. We have unreflectively assumed that all concepts are healthy in a certain sense, and in doing so, we have discriminated against truth. In short, we have treated truth as if it is a normal, healthy concept, when in fact, it is defective and its flaw is inherent. All the paradoxes associated with truth arise from this misunderstanding. As with most bullies, truth's misdeeds are cries for help. However, once we understand its specific defect, we should also recognize that there is no place in our conceptual repertoire for truth. Truth cannot be rehabilitated. Instead, it is time for truth to retire and for us to replace it with one or more healthy concepts that perform its role without causing us trouble. Only by adopting this strategy for handling truth can we finally put an end to its reign of terror.[2]

Accepting that truth is a particular type of defective concept and that we should no longer employ it does not relieve us of our explanatory responsibility. We are still in desperate need of an acceptable theory of truth and an account of the liar paradox, but before we can settle on the best explanation, we need a better grasp of how it has mistreated us. In particular, we need a better understanding of the revenge paradoxes.

I begin by distinguishing between two types of revenge paradoxes: inconsistency problems and self-refutation problems. These problems reinforce one another in the sense that attempts to avoid one tend to bring on the other. In the next section, I argue that if a theory of truth validates the truth rules (i.e. certain intuitively plausible rules governing the use of truth expressions), then either it is restricted from applying to certain languages, which renders it unacceptable, or it faces either an inconsistency problem or a self-refutation problem. Section 13.3 is where I use the revenge paradox phenomenon to justify the theory of truth and the approach to the liar paradox I endorse.[3] I argue that truth displays a particular type of defect: it is an inconsistent concept; roughly, an inconsistent concept has incompatible rules governing the way it should be employed. I present three arguments for theories of truth on which truth is an inconsistent concept. The first argument is an abductive argument: if we accept that truth is an inconsistent concept, then we can explain the pattern of our failures to understand it. That is, the best explanation of why revenge paradoxes occur

[2] In Schiffer's terminology, I am calling for an unhappy face solution to the liar paradox; see Schiffer (2003).

[3] I use both 'theory of truth' and 'approach to the liar paradox' in a loose way to include any set of claims about truth and any set of claims about how to deal with the paradox, respectively.

depends on the claim that truth is an inconsistent concept. The second argument is that if we assume that truth is a consistent concept that obeys the truth rules, then our only options are unacceptable theories. The third argument is that if we accept that truth is an inconsistent concept and we have a proper understanding of how to explain such concepts, then we can construct a theory of truth that: (i) implies that truth obeys the truth rules, (ii) avoids both types of revenge paradoxes, and (iii) does not have to be restricted in any way. Finally, in section 13.4, I present an overview of the theory of truth and the approach I endorse to the liar paradox.

Before I move on to the substantive parts of the chapter, I want to make a methodological point. I am concerned with the liar paradox as it arises in natural languages. Most of the work done by analytic philosophers on the liar paradox focuses on technicalities associated with constructing artificial languages; this is unfortunate. Of course, there is an important place for technical work on truth.[4] However, there is a tendency to get lost in the technical details and ignore how they relate to natural language. For example, one can construct a theory of truth and an artificial language such that the artificial language contains sentences that give rise to the liar paradox, and the theory of truth can handle any sentence that belongs to the language. One can even construct one's theory so that it is expressible in the artificial language without giving rise to revenge paradoxes in that language. That is, we can present an artificial language and a theory of truth for that language such that the theory is expressible in the language and applies to the entire language; hence, we no longer need a substantive object language/metalanguage distinction for certain theories of truth.[5] Although this is a huge accomplishment, I am not concerned with a project of this type because it does not, by itself, constitute an acceptable approach to the liar paradox as it occurs in natural language. To constitute an approach to the liar as it occurs in natural language, a proponent of such a theory would have to claim that natural languages are relevantly similar to the artificial language, or that we should change our natural languages so that they are relevantly similar to the artificial language. It turns out that neither of these claims is tenable because these theories still give rise to revenge paradoxes when applied to anything like natural languages (more on this issue below). Thus, there is more to an acceptable approach to the liar paradox (as it occurs in natural languages) than a theory of truth whose metalanguage is (or is a sublanguage of) its object language. Indeed, it seems to me that the biggest myth associated with contemporary work on the liar paradox is that a theory of truth is 'revenge immune' if and only if it does not require a distinction between

[4] For example, it is important for someone who wants to be able to do mathematical logic in an expressively rich language without worrying about the liar paradox.

[5] See McGee (1991), Field (2003a, 2003b, 2005a, 2005b, forthcoming), and Maudlin (2004) for examples; I am not accusing any of these theorists in particular of getting lost in technical details.

object language and metalanguage.[6] I agree that a theory of truth should apply to the language in which it is formulated, but an acceptable approach to the liar that works for natural languages requires much more than this.

13.2 Revenge Paradoxes

Let us first take a look at the liar paradox.[7] The liar paradox involves sentences like the following (which I call a *liar sentence*):

(1) (1) is false.

The paradox is that from intuitively plausible assumptions via intuitively plausible inferences, one can derive that (1) is both true and false. In fact, there are many different ways to derive this conclusion.[8] The most popular ones depend on T-sentences (i.e. sentences of the form: ⟨p⟩ is true if and only if p)[9], but I prefer one based on what I call the *truth rules*:

(i) *ascending truth rule*: ⟨⟨p⟩ is true⟩ follows from ⟨p⟩.
(ii) *descending truth rule*: ⟨p⟩ follows from ⟨⟨p⟩ is true⟩.
(iii) *substitution rule*: two names that refer to ⟨p⟩ are intersubstitutable in extensional occurrences of ⟨⟨p⟩is true⟩without changing its truth-value.

The argument also depends on some of the inference rules of classical logic. On the one hand, assume that (1) is true. If (1) is true, then '(1) is false' is true (by substitution). If '(1) is false' is true, then (1) is false (by descending). Thus, if (1) is true, then (1) is false. On the other hand, assume that (1) is false. If (1) is false, then '(1) is false' is true (by ascending). If '(1) is false' is true, then (1) is true (by substitution). Thus, if (1) is false, then (1) is true. Therefore, (1) is true if and only if (1) is false. It follows that (1) is

[6] For an extended discussion of this point, see Scharp (TI).

[7] The liar is the most familiar paradox associated with truth, but there are others: the Curry paradox and the Yablo paradox. The Curry paradox is that, from intuitive assumptions, one can use the sentence 'if this sentence is true, then god exists' to derive that god exists (or any other absurdity). The Yablo paradox is that from intuitive assumptions, one can prove contradictory consequences concerning an infinite descending sequence of sentences s_1, s_2, etc., where s_1 is 'for all $i > 1$, s_i is false' and s_2 is 'for all $i > 2$ s_i is false', etc. See van Bentham (1978), Meyer, Routley, and Dunn (1979), Hazen (1990), Beall (1999), Field (2001, 2002, 2003a, 2003b, 2005b) for discussion of the Curry paradox; see Yablo (1985, 1993c), Hardy (1995), Tennant (1995), Priest (1997), Sorenson (1998), Beall (1999, 2001), Leitgeb (2002), Bueno and Colyvan (2003a, 2003b), and Ketland (2004) for discussion of the Yablo paradox.

[8] See Maudlin (2004) for discussion.

[9] '⟨' and '⟩' are angle quotes; 'p' serves as a sentential variable that can be replaced by a sentence, and '⟨p⟩' is the quote-name of such a sentence. I also use 'p' as a logical constant (e.g. in 'p is true'). Note that these uses are distinct: an occurrence of 'p' cannot be both a sentential variable and a constant.

both true and false. Anyone who endorses a theory of truth that applies to a language with sentences like (1) must reject one of the premises, reject one of the inferences, or accept the conclusion.

For most of this chapter, I assume the *principle of mono-aletheism*: no sentence is both true and false; however, I discuss approaches to the liar paradox based on rejecting it in section 3.[10] I also assume that *falsity* is defined in the usual way in terms of truth: the extension of falsity is just the anti-extension of truth, and the anti-extension of falsity is just the extension of truth. Furthermore, I follow most of those who work on the liar paradox in assuming that sentences are primary truth bearers. I should note that the arguments I present do not depend on any particular choice of primary truth bearers.[11]

As I mentioned, a revenge paradox for a theory of truth T often involves a sentence that contains an expression used by T to classify liar sentences. Let us consider an example. Let T be a theory of truth that implies that truth expressions are partially defined predicates. That is, T implies that some sentences containing truth predicates are truth-value gaps—they are neither in the extension of 'true' nor in the anti-extension of 'true'. Assume as well that T validates the truth rules and the other rules involved in the derivation of the liar paradox. Thus, '(1) is true if and only if (1) is false' follows from T. However, no contradiction follows from '(1) is true if and only if (1) is false' because we are working in a three-valued scheme. Indeed, T implies that (1) is a gap. Hence, (1) is not paradoxical for T. We can say that (1) is *pseudo-paradoxical* for T (i.e. (1) has traditionally been involved in a liar paradox, but it poses no problem for T).

So far so good for T; however there is trouble on the horizon. Consider another sentence:

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

Notice that (2) contains 'gap', which is used by T to classify (1). Using the same resources needed to derive '(1) is true if and only if (1) is false', we can derive '(2) is true if and only if (2) is either false or a gap'. On the one hand, assume that (2) is true. If (2) is true, then '(2) is false or a gap' is true (by substitution). If '(2) is false or a gap' is true, then (2) is false or a gap (by descending). Thus, if (2) is true, then (2) is false or

[10] Likewise, I assume the more complex versions of mono-aletheism: no sentence is both true and neither true nor false, and no sentence is both false and neither true nor false.

[11] When I say that sentences are *primary* truth bearers, I mean that I take sentential truth (i.e. the truth of a sentence) to be explanatorily primary; propositional truth, doxastic truth, etc. should be explained in terms of sentential truth. I adopt this view here because the vast majority of those who offer approaches to the liar paradox accept it. Moreover, it seems to me that approaches to the liar that depend on any particular choice of primary truth bearers (e.g. Glanzberg (2004)) face revenge paradoxes of their own.

a gap. On the other hand, assume that (2) is false or a gap. If (2) is false or a gap, then '(2) is false or a gap' is true (by ascending). If '(2) is false or a gap' is true, then (2) is true (by substitution). Thus, if (2) is false or a gap, then (2) is true. Therefore, (2) is true if and only if (2) is false or a gap. Given that (2) is either true, false, or a gap, a contradiction follows. Thus, T is inconsistent. Although T can handle sentences like (1), it cannot handle sentences like (2); (2) constitutes a revenge paradox for T.[12] That is our first example.

Let us consider how T might be altered to accommodate sentences like (2). One way to do so is to alter the logic we use so that the theory still validates the truth rules and still implies that (2) is a gap, but now the theory implies that '(2) is true if and only if (2) is false or a gap' is a gap as well. Let us call this theory T′. Now we cannot derive a contradiction from '(2) is true if and only if (2) is either false or a gap'. However, this sentence poses another problem for T′. Namely, T′ implies that (2) is true if and only if (2) is either false or a gap; hence, T′ has '(2) is true if and only if (2) is either false or a gap' as a consequence. However, T′ implies that '(2) is true if and only if (2) is either false or a gap' is a gap; that is, T′ implies that '(2) is true if and only if (2) is either false or a gap' is neither true nor false. Therefore, T′ implies that one of its consequences is not true. Consequently, T′ is self-refuting—it implies that it is not true.[13, 14] That is our second example.

Let us consider a way of altering the theory so that it is not self-refuting. We need a way of characterizing (2) and '(2) is true if and only if (2) is either false or a gap' that does not result in the theory having a consequence that it labels untrue. One way to do this is to accept that (1) is a truth-value gap, but stipulate that the truth-value gaphood predicate itself is partially defined (i.e. the gaphood predicate has gaps—*gaphood gaps*). Let T″ be such a theory. T″ implies that (1) is a truth-value gap. T″ also implies that (2) is true if and only if (2) is either false or a truth-value gap. However, T″ can be constructed so that it implies that '(2) is true if and only if (2) is false or a truth-value gap' is true; the reason is that T″ does not imply that (2) is either true, false, or a truth-value gap. Indeed T″ implies that (2) is a gaphood gap. Of course, one can construct a new problematic sentence for T″:

(3) (3) is either false, a truth-value gap, or a gaphood gap.

[12] For an example of a theory like T, see Kripke (1975).

[13] For an example of a theory like T′, see Maudlin (2004).

[14] In the last two sentences of this paragraph, I am assuming that one can use 'not' in such a way that 'p is not true' follows from 'p is neither true nor false'. I take it for granted that this is a legitimate use of 'not'. In logic, a sentential operator with this property is often called *exclusion negation*. Thus, I am assuming that sometimes the English word 'not' expresses something like exclusion negation. Note that this assumption need not commit me to the claim that 'not' is ambiguous; see Horn (1989) and Atlas (1989) for discussion. When I am worried about misunderstandings, I use 'Xnot' to express exclusion negation.

However, T″ can follow the same strategy to handle (3) by positing a hierarchy of gaphood predicates, each of which is partially defined. On this account (1) is a truth-value gap, (2) is a gaphood gap, (3) is a gaphood-hood gap, etc. In this way, T″ avoids labeling any of its consequences untrue.[15]

The problem with T″ is that if it applies to a language that contains a completely defined truth-value gaphood predicate, then T″ is inconsistent because it implies that a sentence of this language like (2) (i.e. a sentence that attributes either falsity or truth-value gaphood to itself—where the truth-value gaphood predicate is completely defined) is true if and only if it is either false or a truth-value gap. Thus, T″ faces an inconsistency problem. The progression from T to T′ to T″ illustrates the fact that there is something like an oscillation between the two kinds of revenge paradoxes—attempts to avoid one tend to bring on the other.[16]

It is my view that one must distinguish between these two types of revenge paradoxes in order to understand our current predicament regarding truth. In short, there are two broad trends when it comes to theories of truth designed to handle sentences like (1). Some theories can handle sentences like (1), but they still have inconsistent consequences for other sentences (e.g. (2)). That is, they do not provide a way of solving all other paradoxes associated with truth that are structurally identical to the liar. This is the inconsistency problem. Other theories can handle sentences like (1), but they imply that they have the same status (i.e. being untrue) as (1). Because few, if any, theories of truth that are designed to handle sentences like (1) imply that sentences like (1) are true, a theory of truth that implies that it has the same status as a liar sentence implies that it is untrue. This is the self-refutation problem.[17]

The inconsistency problem arises when a theory of truth handles some versions of the liar paradox, but not all of them. There are many different versions of the liar; some versions involve concepts that are often used to classify sentences that figure in other versions. This should not come as a surprise given the prominence of views on which sentences that figure in liar paradoxes are defective in a way that renders them neither true nor false. Once one has a term for the third status, one has a new version of the liar paradox. The most common response to the inconsistency problem is to restrict the theory so that it does not apply to such sentences.

[15] For an example of a theory like T″, see Field (2003a, 2003b, 2005a, 2005b, forthcoming).

[16] I borrow the term 'oscillation' from McDowell (1994).

[17] Both the Curry paradox and the Yablo paradox depend on the truth rules as well (i.e. they require all three rules for their construction), and approaches to each one generate revenge paradoxes in the same way that approaches to the liar paradox generate revenge paradoxes; thus, one can use structurally analogous arguments to the ones in this chapter to argue for conclusions pertaining to the Curry and the Yablo that are analogous to the conclusions I draw pertaining to the liar. It is my view that all three paradoxes (i.e. the liar, the Curry, and the Yablo) are manifestations of the defectiveness of our concept of truth. The approach to truth that I offer solves all three without generating revenge paradoxes of any kind.

On the other hand, the self-refutation problem arises in connection with the consequences of a theory of truth. The liar paradox is unlike other paradoxes (e.g. Russell's paradox, Grelling's paradox, etc.) in that it concerns truth, which applies to things that can participate in inferential relations (e.g. sentences, propositions, etc.). In other cases, the paradoxical items (e.g. sets, predicates, etc.) are not the type of thing that can be the consequence of a theory. However, for truth, the paradoxical items are sentences, which can be consequences of a theory. A theory of truth that is designed to deal with the liar paradox has to classify many paradoxical sentences like (1). It turns out that for many theories of truth, no matter what they say about such sentences, some of these sentences are going to be consequences of the theory.

This way of formulating the self-refutation problem is somewhat misleading because it makes it seem as though the class of paradoxical sentences is fixed. In fact, we should think of paradoxicality as relative to a theory of truth, and we should distinguish between paradoxicality and pseudo-paradoxicality. A sentence is *paradoxical* for a theory of truth if and only if the theory of truth either has contradictory consequences for it or has consequences for it that the theory implies are untrue. A sentence is *pseudo-paradoxical* for a theory of truth if and only if it is not paradoxical for the theory in question, but it is paradoxical for the naïve theory of truth, which implies that truth is completely defined and obeys all the principles we commonly take truth to obey (e.g. the truth rules, rules about how it interacts with sentential operators, etc.). The sentences that are pseudo-paradoxical for a theory of truth (e.g. (1)) figure in the liar paradox, but the theory can handle them. The sentences that are paradoxical for a theory of truth figure in revenge paradoxes for the theory, which the theory cannot handle. The class of sentences that are pseudo-paradoxical for a theory of truth and the class of sentences that are paradoxical for a theory of truth depend on the way the theory of truth classifies the liar (e.g. as gappy, as indeterminate, as uncategorical, etc.). For example, the revision theory of truth implies that the liar is uncategorical. Thus, a revenge paradox for it concerns the sentence:

(4) (4) is either false or uncategorical.

That is, sentence (4) is paradoxical for the revision theory of truth. However, an indeterminacy theory of truth (i.e. one on which the liar is indeterminate) has no problem with (4) because it does not imply that the liar is uncategorical. Thus, (4) is not paradoxical for an indeterminacy theory of truth. It turns out that if a theory of truth validates the truth rules, then no matter what it says about the liar sentence, the set of sentences that are pseudo-paradoxical for it will include some of its own consequences (unless it is restricted so that it does not apply to them). Given that an acceptable theory of truth does not imply that pseudo-paradoxical sentences are true, one can either restrict one's theory so that it does not apply to the pseudo-paradoxical

sentences that are outside its scope, or one can bite the bullet and accept that one's theory implies that some of its consequences are untrue.

13.3 The Revenge Argument

In this section, I present a criticism of theories of truth that offer approaches to the liar paradox on which truth is pretty much as we take it to be. In particular, it is a criticism of theories of truth that validate the truth rules (i.e. theories of truth that imply that the truth rules are valid for some class of sentences that includes some liar sentences). I argue that any theory of truth that implies that the truth rules are valid is either: (i) inconsistent, (ii) self-refuting, or (iii) restricted so that it does not apply to certain sentences that contain truth predicates. Although there are many theories of truth that offer approaches to the liar paradox on which one or more of the truth rules are not valid, I do not address them here.[18] It is my view that the truth rules are constitutive of our concept of truth—any theory of truth that implies that truth does not obey them is unacceptable. Of course, that claim is not intended to be a criticism. However, one can develop it into a criticism that shows these theories to be unacceptable, but I don't have the space to do so here.[19]

Assume that T is a theory of truth and T implies that the truth rules are valid for a class of sentences that includes (1). Assume also that T implies that truth predicates are univocal, invariant, non-circular, etc.; in short, truth predicates do not have any 'hidden' semantic features that render the reasoning in the liar paradox invalid. Note that this assumption does not add much because theories of truth that imply that truth predicates have 'hidden' semantic features don't usually validate the truth rules (at least, I am unaware of any that do).[20] Finally, assume that T applies to a language that contains liar sentences. Let a *liar sentence* be any sentence that attributes falsity and only falsity to itself. Thus, sentence (1), 'this sentence is false', and 'the sentence named by the third singular term used in the sixth sentence of the second paragraph of the second section of 'Aletheic Vengeance' is false', are liar sentences.

There is plenty to say about what languages and sentences are, and about what it is for a theory to apply to a particular language or to a particular sentence, but I want to

[18] e.g. theories of truth that treat natural language truth predicates as context-dependent (e.g. Parsons (1974), Burge (1979), Barwise and Etchemendy (1987), Gaifman (1992, 2000), Koons (1992, 2000), Simmons (1993), and Glanzberg (2004)), theories of truth that treat truth as a circular concept (e.g. Herzberger (1982a, 1982b), and Gupta and Belnap (1993)), theories of truth that reject the substitution rule (e.g. Skyrms (1982), and theories of truth that reject the ascending rule (e.g. Feferman (1982)).

[19] See Scharp (AP) for this criticism.

[20] For examples of theories of truth that imply that truth predicates have 'hidden' semantic features, see Burge (1979), Gupta and Belnap (1993), and Williamson (2000).

leave them at an intuitive level. In addition, there is plenty to say about the conditions under which a language contains a liar sentence; however, it is my view that because of the prevalence of *empirically paradoxical sentences* (i.e. sentences that are paradoxical because of some empirical facts) and *inter-linguistic truth attributions* (i.e. sentences of one language that attribute truth or falsity to sentences of other languages) it is impossible to provide a non-circular account of the conditions under which such sentences arise in natural languages.[21]

There are very few choices for the way in which a theory of truth that validates the truth rules classifies a liar sentence. The theory can imply that liar sentences are false or the theory can imply that liar sentences are true, but given that the theory implies that the truth rules are valid, a theory of either type implies that liar sentences are true if and only if they are false; for a theory that classifies the liar as true or false, '(1) is true if and only if (1) is false' is a contradiction. Therefore, an acceptable theory of truth that validates the truth rules will not classify liar sentences as true or false.

Instead of classifying (1) as true or as false, T can classify (1) as neither true nor false. We English speakers find this a natural description of the case. That is, we find it natural to say that certain things are neither true nor false (e.g. acorns). We also find it natural to say that such things are not true and not false. Here we are using 'not', but this use is distinct from the use of 'not' in 'a sentence that is not true is false'; the former 'not' expresses exclusion negation, while the latter expresses choice negation.[22] In English, we sometimes use 'not' to express exclusion negation as above. Other times we use 'not' to express choice negation. I assume that it can express exclusion negation and that it can express choice negation.

If T implies that the liar is a truth-value gap, then we can construct another sentence that causes problems for T. Sentence (2) (i.e. '(2) is either false or a gap') is classified as a gap by T and, hence, it is a consequence of T; that is, if T implies that (2) is a gap, then T implies that (2) is either false or a gap. Thus, T has (2) as a consequence and T implies that (2) is untrue.[23] Thus, T faces a self-refutation problem.

In addition, T faces an inconsistency problem. Recall that if T validates the truth rules, then '(1) is true if and only if (1) is false' is a consequence of T. The approach to the liar on which truth is treated as a partially defined concept (i.e. on which (1) is a gap) handles (1) because both (1) and '(1) is true if and only if (1) is false' are gaps. The catch is that '(1) is true if and only if (1) is false' is not a genuine contradiction according to this theory; that is, one cannot prove '(1) is true and (1) is not true' from '(1) is true if and only if (1) is false' because the proof depends on the principle of bivalence for (1) (i.e. (1) is either true or false), which the partiality approach denies.

[21] On the former, see Scharp (RB) and on the latter see Scharp (TI) and Eklund (forthcoming).

[22] See n. 14.

[23] This argument depends on the standard rule of disjunction introduction.

However, the partiality approach cannot rely on the same trick when it comes to sentence (2). The sentence '(2) is true if and only if (2) is either false or a gap' follows from T by an argument that is structurally identical to the one that shows '(1) is true if and only if (1) is false' follows from T. Given that a sentence is either true, false, or a gap, one can derive that (2) is both true and either false or a gap from this consequence. Hence, '(2) is true if and only if (2) is either false or a gap' is a consequence of T and it is a genuine contradiction. Thus, T faces an inconsistency problem.

In summary, if T validates the truth rules and T implies that liar sentences have status Δ, where a sentence is Δ only if it is Xnot true, then there are three options for T:

(i) T implies that $(2')$, '$(2')$ is either false or Δ', is true.
(ii) T implies that $(2')$ is false.
(iii) T implies that $(2')$ is Δ.

On any of these options, T implies '$(2')$ is true if and only if $(2')$ is false or Δ'. If T classifies this sentence as true, then T is inconsistent. If T classifies it as false, then T is self-refuting. If T classifies it as gappy, then T is self-refuting. Therefore, T is either inconsistent or self-refuting.

Given the massive amount of work on the liar paradox and the ridiculously sophisticated logical tools that have been marshaled to combat it, the reader *should* be skeptical when presented with such a simple argument that is touted as a refutation of most prima-facie plausible approaches to the liar paradox. Although the distinction between the inconsistency problem and the self-refutation problem is new, the revenge argument should not come as a surprise to any of the veterans of our battles with the liar paradox. In fact, most of them have been hard at work devising plans to avoid arguments like this one. So why have I presented it as a central insight into the nature of truth and the liar paradox? It might seem like I am throwing a rock at an army battalion; if so, read on—it turns out that the appearance of a battalion is nothing but a mirage and a thrown rock is a fine way to expose it as such. That is, the real insight is that there is no acceptable way of avoiding the revenge argument; thus, the real work is done in the objections and replies, to which I now turn.

Objection 1: One can avoid the liar paradox and both types of revenge paradoxes by assuming that paradoxical sentences are meaningless or ill-formed.

Reply 1: Strictly speaking, this is not an objection to the revenge argument—it offers an approach to the liar that would take care of the liar and all the revenge paradoxes. However, it is instructive to see why this sort of approach fails. The most obvious problem with this objection is that there is no independent reason to think that paradoxical sentences are meaningless or ungrammatical. In fact, if one were to adopt such an account, one would have to reject our most popular theories of meaningfulness and theories of grammar. There is another reason to reject these accounts: paradoxicality is not determined by meaning and grammar.

That is, one can specify two sentence tokens of the same type that have the same sentential meanings, the same subsentential meanings for their subsentential parts, and the same referents for their singular terms, but one is paradoxical and the other is not. Paradoxicality can depend on virtually any fact one can imagine, while meaningfulness and grammaticality do not. Thus, if one accepts that paradoxical sentences are meaningless or ungrammatical, then one has to accept that whether a sentence is meaningful or grammatical can depend on virtually any fact that one can imagine, which is radically implausible.[24]

Objection 2: The revenge argument involves not just inferences licensed by the truth rules, but inferences of classical logic as well. If one endorses a non-classical logic as part of one's approach to the liar paradox, then one can avoid both types of revenge paradoxes.

Reply 2: I do not deny that many approaches to the liar paradox involve non-classical logics. However, using this move to block the revenge argument has several problems. One problem is that the classical inference rules needed to derive the troublesome conclusion (i.e. '(2) is true if and only if (2) is either false or gappy') are minimal. All one really needs is a conditional that obeys the natural deduction inference rule of conditional proof (alternatively, a conditional for which one can prove a deduction theorem); a conditional with this property is required if anything like everyday reasoning is possible in the language. These inference rules are going to be valid for any natural language; thus, a theory of truth that rejects them will not apply to natural languages.

Moreover, a theory of truth that avoids the revenge argument by denying one of the inference rules involved (except for the truth rules of course) would still have to be restricted so that it does not apply to languages for which the inference rule in question is valid. Let T be a theory of truth that applies only to languages for which a certain inference rule R involved in the revenge argument is invalid. Let L be a language for which R is valid and let L contain a truth expression. Of course, T does not apply to L; thus, T is restricted so that it does not apply to some languages that contain truth expressions. A theory of truth that avoids the liar paradox or the revenge paradoxes only by denying certain inference rules of classical logic is a theory that is restricted so that it does not apply to certain languages containing truth expressions. Thus, the conclusion of the revenge argument (i.e. a theory of truth that validates the truth rules is either inconsistent, self-refuting, or restricted) withstands this objection.[25]

[24] See Kripke (1975) for a similar point; see also Scharp (RB) for discussion.

[25] I suppose that one could deny that such languages exist, but it would follow that there is no language such that it is one in which we can reason normally and it contains a truth expression; that seems radically counterintuitive to me.

Objection 3: Instead of treating the expression for truth-value gaps as completely defined, which is an assumption needed to derive the inconsistency problem in the revenge argument, one can assume that 'gappy' is itself gappy. Then one can treat (1) as a truth-value gap and one can treat (2) as a gaphood gap. Indeed, one can define a hierarchy of partially defined gaphood predicates that can be used to classify all the sentences that seem to give rise to liar paradoxes, and one can do so without facing either a self-refutation problem or an inconsistency problem.

Reply 3: I agree that one can provide a theory of truth of this sort and an artificial language with predicates like these such that none of the sentences of the language give rise to revenge paradoxes for the theory.[26] However, that does not constitute an acceptable approach to the liar paradox because the theory has to be restricted so that it does not apply to languages that contain completely defined gaphood predicates; otherwise, sentences like (2) in which completely defined gaphood predicates occur are paradoxical for the theory (I discussed this problem in section 1).

The objector might respond by claiming that if there were a language that obeys the logic posited by the theory of truth in question and contains a truth predicate and a completely defined indeterminacy predicate, then it would be trivial in the sense that anything would be derivable in it; thus, if the theory is right, then languages with truth predicates don't have completely defined indeterminacy predicates. My reply is that we can just stipulate that some language contains a truth predicate and a completely defined indeterminacy predicate; thus, the theory of truth in question won't apply to this language because whatever the right logic is for the language, it won't be the one posited by the theory. It is common to assume that we can stipulate the syntactic features of a language and that we can stipulate that a certain word expresses a certain concept. Thus, it makes sense to think that we can stipulate that a language has a truth expression and a completely defined indeterminacy predicate.[27]

Objection 4: One can arrive at a satisfactory theory of truth by restricting it so that it does not apply to sentences that give rise to inconsistency problems or self-refutation problems. The most familiar strategy of this type is to assume that the theory is formulated in one language (the metalanguage) and applies only to languages that are expressively weaker in certain ways (the object languages). A theory of this type is restricted from applying to languages that have the expressive resources required to formulate the theory. However, some more recent theories of truth do not appeal to the distinction between object language and metalanguage, but they still have to be restricted to avoid revenge paradoxes. For example, Field's theory of truth applies

[26] See Field (2003a, 2003b, 2005a, 2005b, forthcoming) for an example of such a theory.

[27] Some philosophers deny that we have this stipulative power, but I cannot take issue with them here; see Williamson (1997).

unproblematically to certain artificial languages that have the resources to formulate the theory; however, it does not apply to languages that contain completely defined indeterminacy predicates (the linguistic expressions that give rise to revenge paradoxes for Field's theory—completely defined indeterminacy predicates—aren't required to formulate the theory).

Reply 4: Although almost everyone who presents an approach to the liar resorts to this move in one way or another, it is unacceptable because it results in a theory of truth that does not even apply to all truth expressions. Desperation has overwhelmed common sense in this case. It is as if these philosophers are saying 'look everyone, I have come up with a theory of chairs!' When a critic objects, 'your theory implies that *this* chair is both black and not black—that result refutes your theory', the theorist responds, 'oh, my theory doesn't apply to *that* chair'. We all *should* agree that this response is totally unacceptable. It often comes as a shock to those outside philosophical logic that this sort of move is tolerated for truth theorists who offer approaches to the liar.

It seems to me that this response to theories of truth that are restricted to avoid revenge paradoxes should be sufficient. However, I expect that those philosophers who have been hardened by combat with the liar will be deaf to this sort of criticism. In a companion paper, I argue that theories of truth that have been restricted to avoid revenge paradoxes are unacceptable.[28] There I argue that if T is a theory of truth that is restricted to avoid revenge paradoxes and L is a natural language, then there are sentences of L that give rise to revenge paradoxes for T; thus, if T applies to these sentences of L, then T is either inconsistent or self-refuting. The key to the argument is constructing a sentence of L that attributes truth indirectly to a sentence of some other language that gives rise to a revenge paradox for L; that is, if a theory of truth faces revenge paradoxes, then one can 'import' one of these revenge paradoxes into a natural language. It is important that one can construct such a sentence even if one can appeal only to language-specific concepts of truth (e.g. truth-in-L, truth-in-English). Therefore, if a theory of truth is restricted to avoid revenge paradoxes, then it does not successfully apply to natural languages.

Objection 5: All the linguistic expressions used to construct revenge paradoxes are meaningless. The newest generation of approaches to the liar paradox show that one can construct a theory of truth that applies to the language in which it is formulated and classifies all the sentences of that language without giving rise to revenge paradoxes.[29] Of course, these theories seem to face revenge paradoxes when applied to languages containing other linguistic resources. However, one can treat

[28] See Scharp (IT).

[29] See McGee (1991), Field (2002, 2003a, 2003b, 2004, 2005a, 2005b, forthcoming), and Maudlin (2004) for examples.

these linguistic expressions as meaningless and avoid the revenge paradoxes altogether. If one takes this path, then one does not even have to restrict one's theory of truth.

Reply 5: Some philosophers do try to avoid restricting their theories of truth by claiming that the resources that give rise to the revenge paradoxes are meaningless or unintelligible.[30] I call this the *unintelligibility maneuver*. My view is that it is unacceptable to assume that these linguistic expressions are meaningless. As I have said, for a language that contains truth-value gaps, one can define two sentential operators that behave like classical negation:

p	$\sim$p	$\neg$p
T	F	F
F	T	T
G	G	T

The first one ('$\sim$') is choice negation and the second ('$\neg$') is exclusion negation. Both can be expressed in English.[31] Even if we assume that some sentences are neither true nor false, one might infer from the claim that a sentence is not true that the sentence is false. If this inference is appropriate, then the negation involved is choice negation. On the other hand, one might infer from the claim that a sentence is neither true nor false that the sentence is not true. If this inference is appropriate, then the negation involved in the conclusion is exclusion negation. A theory like McGee's or Field's or Maudlin's has trouble with sentences that express exclusion negation for two reasons. First, these theories are fixed-point theories and so apply only to languages that do not contain non-monotonic sentential operators.[32] However, exclusion negation is non-monotonic. Thus, they do not even return results for languages that express exclusion negation. Second, one can easily extrapolate to determine the results they would return if they were capable of returning results. The sentence

(5) (5) is Xnot true.

poses a problem. It means something like *(5) has a status other than that of being true*. One can use (5) to generate a revenge paradox for most theories of truth that imply that (5) is

[30] See Parsons (1984), Priest (1990), and Tappenden (1999), who claim that there is no such thing as exclusion negation; see also Maudlin (2004), who claims that there are no non-monotonic sentential operators whatsoever.

[31] For evidence of this claim see Atlas (1989) and Horn (1989).

[32] In a three-valued scheme, a sentential operator is *monotonic* if and only if for a sentence containing that sentential operator, changing a component of that sentence from a gap to a truth-value (i.e. from a gap to true or from a gap to false) never results in changing the sentence from one truth-value to the other or from a truth-value to a gap (i.e. from true to false, from false to true, from true to a gap, or from false to a gap). Intuitively, one can 'fill in' the gaps in the components without changing the truth-value of the compound. See Gupta and Martin (1984), who show that, by using a weak Kleene scheme, one can arrive at fixed points even though one's language contains certain non-monotonic operators. However, exclusion negation is not among them.

Xnot true. Likewise, one can generate revenge paradoxes using completely defined gaphood predicates (as I did in the revenge argument), paradoxicality predicates, groundedness predicates, certain conditionals, quantification over hierarchies of predicates, and so on. In order to pursue the strategy advocated in the objection, one would have to claim that all these linguistic items are meaningless.

I take it for granted that if there is an established practice of using a linguistic expression, then that linguistic expression is meaningful.[33] For each of the linguistic expressions that are labeled unintelligible by these theorists, there is an established practice of using them. Moreover, these linguistic expressions belong to some natural languages, including English. Furthermore, anyone who claims that the linguistic expressions involved in revenge paradoxes are meaningless will have to claim that logicians and linguists have been wasting their time studying exclusion negation, other non-monotonic sentential operators, and the rest of the 'outlaw' linguistic expressions.

In addition, the 'outlaw' linguistic expressions serve an important explanatory role. If we decided that they are all meaningless and gave them up, then we would rob natural languages of important expressive resources. For example, if an object, A, is in neither the extension nor the anti-extension of a predicate, ϕ, then we need a way of expressing this fact. One way of doing so is to say that A is Xnot ϕ and A is Xnot $\lfloor\sim \phi\rfloor$. Another is to say that A is a ϕ-value gap. If the theorists in question are right that the 'outlaw' linguistic expressions are meaningless, then we have no way of expressing these facts.

Finally, simply claiming that the linguistic resources in question are meaningless is not enough to avoid the revenge argument. One would have to provide an independent argument for this claim (e.g. something other than, 'that's the only way to avoid the liar paradox'). No such argument has been forthcoming and it seems it would be impossible to present one whose premises were more plausible than the claim that these items are meaningful.[34]

Analytic philosophy has a long history of claiming that certain linguistic expressions that figure in established linguistic practices are meaningless. It is high time for us to realize that we need no longer resort to this kind of move; we can provide an approach to the liar paradox without it.

Objection 6: Theories of truth that are restricted to avoid revenge paradoxes can be thought of as revisionary, not descriptive. That is, one can treat them as prescribing

[33] Even linguistic expressions like 'tonk' are meaningful; they just express inconsistent concepts; see Prior (1960) for a discussion of 'tonk'. Of course, Wittgenstein (1923) is infamous for claiming that many seemingly meaningful words are 'nonsense' (*unsinnig*); I don't have the space to discuss the relation between his views on language and the claim on which this footnote comments.

[34] See Eklund (forthcoming) for another criticism of the unintelligibility maneuver that focuses on exclusion negation.

how we should use truth predicates instead of describing how we do use them. If one treats a restricted theory in this way, then a proponent of such a theory advocates eliminating from natural language the resources that contribute to revenge paradoxes (e.g. non-monotonic sentential operators, gapless gaphood predicates, etc.).

Reply 6: I think it is fine to treat these theories as revisionary, and there is an important place for such theories.[35] However, qua revisionary theories of truth, they face several problems. First, they say nothing about natural languages as they are now. Thus, they do not really provide approaches to the liar paradox at all. Second, even as revisionary theories of truth, they fail. As long as one has a truth predicate in the language that obeys the truth rules, one can 'import' the revenge paradoxes into the language even though the language does not have the 'outlaw' linguistic resources. Assume that T is a theory of truth that offers an approach to the liar paradox and that T is a revisionary theory of truth—it implies that we should change English so that the linguistic items involved in revenge paradoxes (e.g. exclusion negation, other non-monotonic operators, gapless gaphood predicates, etc.) are no longer part of English. Call the new language English*. Assume also that T validates the truth rules. I am willing to grant that T might be expressible English* and that T can adequately classify all the sentences of English* that involve truth attributions to sentences of English*.[36] The problem arises with sentences of English* that attribute truth to sentences of other languages. Given that the revenge argument is correct (i.e. sound), T faces revenge paradoxes when applied to other languages. Let L be a language that contains the resources needed to construct a revenge paradox for T. Although T is restricted so that it does not apply to L, we know that if it did apply to L, then it would be inconsistent or self-refuting because L has sentences that would give rise to revenge paradoxes for T. Call such sentences *potentially paradoxical for T*. I argue that English* contains potentially paradoxical sentences for T; hence, to avoid inconsistency or self-refutation, T has to be restricted so that it does not apply to certain sentences of English*. Moreover, there is no a priori way to specify which sentences of English* are potentially paradoxical for T. Thus, even though English* does not contain any of the explicitly problematic linguistic resources, T is unable to apply to all the sentences of English*.

Consider the inconsistency case. Let ϕ be a sentence of L that is potentially paradoxical for T, let the potential paradox be an inconsistency problem, and let English* have a name 'ϕ' for ϕ. Let ψ be the English* sentence, 'ϕ is true'. If ϕ figures in an inconsistency problem for T, then ψ figures in an inconsistency problem for T.

[35] e.g. these theories can prescribe the parts of language that are acceptable for formulating arguments in mathematical logic.

[36] Field's theory is a good example of a theory like T. Others include Maudlin's and McGee's.

Therefore, if T does not face an inconsistency problem, then it is restricted so that it does not apply to ψ.

In order to justify the claim that if ϕ is potentially paradoxical for T, then ψ is too, we need to consider an example (remember that whether a sentence is paradoxical for a theory of truth depends on the features of the theory). Assume that T is a version of Kripke's theory of truth. A revenge liar for Kripke's theory is:

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

Assume that (2) is a sentence of L, that English* does not contain sentences like (2) because English* does not have a gaphood predicate, and that T is restricted so that it does not apply to L. If T did apply to L, then T would imply that (2) is both true and either false or a gap. However, as long as English* contains a name, '(2)' for (2), English* contains:

(6) (2) is true.

If T applies to (6), then T implies that (6) is both true and either false or a gap. Assume that (6) is true. If (6) is true, then '(2) is true' is true (substitution). If '(2) is true' is true, then (2) is true (descending). If (2) is true, then '(2) is either false or a gap' is true (substitution). If '(2) is either false or a gap' is true, then (2) is either false or a gap (descending). If (2) is either false or a gap, then '(2) is true' is either false or a gap (ascending).[37] If '(2) is true' is either false or a gap, then (6) is either false or a gap (substitution). Hence, if (6) is true, then (6) is either false or a gap. Assume that (6) is either false or a gap. If (6) is either false or a gap, then '(2) is true' is either false or a gap (substitution). If '(2) is true' is either false or a gap, then (2) is either false or

[37] It might seem that this inference is not licensed by the truth rules and instead requires a rule specific to weak truth. In three-valued settings, it is common to distinguish between weak truth and strong truth, where *weak truth* obeys the rule 'if p is gappy, then "p is true" is gappy', but *strong truth* obeys the rule 'if p is gappy, then 'p is true' is false'. As far as I know, no one has noticed that one can justify the weak truth rule by appeal to the truth rules and the claim that they can be applied not only to an entire formula, but to the subformulas of a formula. We can represent the weak truth rule as: from ⟨⟨p⟩ is gappy⟩ infer ⟨⟨p⟩ is true⟩ is gappy⟩. Given our definition of the gaphood predicate, this rule becomes: from ⟨¬⟨⟨p⟩ is true⟩ ∧ ¬ ∼ ⟨⟨p⟩ is true⟩⟩ infer ⟨¬⟨⟨⟨p⟩ is true⟩ is true⟩ ∧ ¬ ∼ ⟨⟨⟨p⟩ is true⟩ is true⟩⟩. We can derive this rule using the ascending truth rule on subformulas:

1. ¬⟨⟨p⟩ is true⟩ ∧ ¬ ∼ ⟨⟨ p⟩ is true⟩
2. ¬⟨⟨⟨p⟩ is true⟩ is true⟩ ∧ ¬ ∼ ⟨⟨p⟩ is true⟩
3. ¬⟨⟨⟨p⟩ is true⟩ is true⟩ ∧ ¬ ∼ ⟨⟨⟨p⟩ is true⟩ is true⟩

It is common practice to apply an inference rule to subformulas so long as it is an equivalence rule. Given that I have endorsed both the ascending and descending truth rules, we have essentially an equivalence rule. There are plenty of issues associated with this move (e.g. freestanding vs. ingredient content, categorical vs. hypothetical inference rules, etc.), but I do not have the space to discuss them.

a gap (descending).[38] If (2) is either false or a gap, then '(2) is either false or a gap' is true (ascending). If '(2) is either false or a gap' is true, then (2) is true (substitution). If (2) is true, then '(2) is true' is true (ascending). If '(2) is true' is true, then (6) is true (substitution). Hence, if (6) is either false or a gap, then (6) is true. Therefore, (6) is true if and only if (6) is either false or a gap. It follows that (6) is both true and either false or a gap. The self-refutation case is analogous.[39, 40]

One might stipulate that English* does not have names for sentences of L, so this way of 'importing' revenge paradoxical sentences into English* doesn't work. The problem with this response to the argument is that there are plenty of ways English* might have to refer to sentences that are potentially paradoxical for T. For example, English* might have a definite description that picks out a potentially paradoxical sentence for T, a sentence of English* containing a demonstrative might refer to a potentially paradoxical sentence for T, a pronoun of English* might refer to a potentially paradoxical sentence for T, a quantifier of English might range over a potentially paradoxical sentence for T, etc. There is no way to specify which of English*'s linguistic resources might refer to a potentially paradoxical sentence of T because it might depend on all sorts of empirical facts (e.g. 'the last sentence Russell uttered is true' might be potentially paradoxical for T if the empirical facts turn out unfavorably). In addition, even if English* has no way to refer directly to a potentially paradoxical sentence for T, we can still 'import' a revenge paradox into English* by considering a sentence of English* that attributes truth to a sentence ϕ of some language L, where ϕ attributes truth to a sentence θ, which is potentially paradoxical for T. If one wants to ensure that no sentence of English* indirectly attributes truth to a sentence that is paradoxical for T, then one has to eliminate from English most of the linguistic resources we assume natural languages have. If instead one wants to ensure that T does not apply to any potentially paradoxical sentences for T, then one has to restrict T so that it does not apply to most of the sentences of English* that contain truth predicates. Either way, T fails to offer an acceptable approach to the liar paradox.

Objection 7: Some of the replies I have offered (i.e. 2 and 6) depend on inter-linguistic truth attributions (i.e. sentences of the form 'p is true' where 'p' names a sentence of some other language). There are two problems with this move. First, all the theorists being discussed provide theories for language-specific truth predicates. A *language-specific truth predicate* (an *LS truth predicate*) is satisfied only by true sentences of

[38] Again, this inference requires using the descending truth rule on a subformula.

[39] See Eklund (forthcoming) for a different sort of 'importation' argument.

[40] This argument highlights the fact that if a truth expression obeys the truth rules and applies to sentences of languages other than the one to which it belongs, then applying the descending truth rule can take one from a sentence of one language to a sentence of another. Although this consequence doesn't seem problematic to me, I don't have the space to defend it here.

a particular language. For example, 'true-in-English' is an LS truth predicate: 'p is true-in-English' is true if p is a true sentence of English, and it is false if either p is a false sentence of English or p is a sentence of some other language. Most of these theorists claim that natural language truth predicates can be explained in terms of LS truth predicates. The arguments given so far fail for LS truth predicates and, hence, for natural language truth predicates (as long as we explain the latter in terms of the former). For example, if T is a theory of truth-in-L for a particular language L, and T offers a partiality approach to the liar paradox, then one might think that T would face a revenge paradox from the following sentence of some other language, L:

(2″) (2″) is either false-in-L or gappy-in-L.

However, the reasoning employed in the revenge argument to derive either an inconsistency problem or a self-refutation problem fails for (2″). Indeed (2″) is either false-in-L or gappy-in-L.[41] Therefore, none of the arguments offered so far even threatens the theories of truth that have been discussed.

Second, I have assumed both that sentences are primary truth bearers and that natural language truth predicates are unrestricted (i.e. that they are not LS truth predicates). However, these two commitments are incompatible. Consider a sentence that is a member of English and German (e.g. in English, 'Kripke rang' means that Kripke rang, and in German, it means that Kripke wrestled). If the truth conditions for the sentence qua member of English are satisfied and the truth conditions qua member of German are not, then the sentence is both true and false (i.e. it is true qua member of English and false qua member of German). Of course, no sentence can be both true and false. If I want to keep the claim that sentences are truth bearers, then I have to admit that natural language truth predicates are language-specific. It seems that I do not even grasp the rules of the game.

Reply 7: I have three replies to this objection. First, it is acceptable to claim that natural language truth predicates are unrestricted and choose sentences as primary truth bearers as long as one is careful in specifying what sentences are. It is common to distinguish sentence tokens from sentence types; *sentences tokens* are physical objects while *sentence types* are abstract entities. Beyond this claim, there is little agreement on how to individuate sentence tokens, how to individuate sentence types, how to explain sentence types, or how to explain the relation between sentence tokens and sentence types. One option is to take sentence tokens to be pairs of possible physical objects and contexts, where the context is sufficient to determine the syntactic and semantic features of the sentence token. One can then take 'true' to express an unrestricted notion of truth and treat sentence tokens as primary truth bearers without running

[41] Of course, '(2″) is true-in-L' does not follow from '(2″) is false-in-L or gappy-in-L' and the fact that (2″) is '(2″) is false-in-L or gappy-in-L'; for that result, (2″) would have to be a sentence of L.

into trouble. Another option is to take sentence types to be individuated on the basis of semantic features. One can then take 'true' to express an unrestricted notion of truth and treat sentence types as primary truth bearers without difficulty. There are many other options as well.[42]

Second, I agree that almost everyone who presents a theory of truth that offers an approach to the liar paradox addresses only LS truth predicates and assumes that natural language truth predicates can be explained in terms of LS truth predicates. However, natural language truth predicates cannot be explained in terms of LS truth predicates; of course, there is no way I can provide a good justification for this claim here. I do want to point out several problems with it. One is that those who claim that natural language truth predicates can be explained in terms of LS truth predicates rarely say how this is to be done. Those who do usually either claim that natural language truth predicates are ambiguous and can take on the meaning of any of the LS truth predicates or that a truth predicate of a natural language L is synonymous with 'translatable into a sentence of L that is true-in-L'. Neither of these approaches is plausible. First, they imply that some warranted assertions of blind truth attributions are unwarranted (e.g. Cletus says that some sentence is true without knowing the language to which it belongs).[43] Second, they cannot handle multiple-target truth attributions (e.g. when Cletus says 'everything Brandine uttered yesterday is true' and Brandine spoke in several different languages on the day in question).[44] Third, the ambiguity approach faces the problem that 'true' fails all the standard ambiguity tests used by linguists, and the translation approach cannot handle truth attributions to some sentences of other languages that contain LS truth predicates (e.g. 'Schnee ist weiss' ist wahr-auf-Deutsch' isn't translatable into English unless English has 'true-in-German'; thus, a translation predicate and 'true-in-English' aren't sufficient).[45] Finally, they do not validate the truth rules (e.g. the ascending truth rule for an LS truth predicate is: ⟨⟨p⟩ is true-in-L⟩ follows from ⟨p⟩ and ⟨⟨p⟩ is in L⟩, but not from ⟨p⟩ alone).

Third, all the arguments offered so far that depend on inter-linguistic truth attributions can be altered to accommodate the claim that natural language truth predicates can be explained in terms of LS truth predicates. For example, even if

[42] See Kaplan (1973, 1990), Hugly and Sayward (1981), Simons (1982), Bromberger (1989), Wetzel (1993), Horwich (1998: 98–103), Cappelen (1999), Dummett (1999), Szabó (1999), and Truncellito (2000) for discussion of the distinction between types and tokens. I prefer treating sentence tokens (i.e. pairs of possible objects and contexts) as primary truth bearers, so long as one individuates contexts finely enough. However, there are some tricky issues here having to do with the fact that the syntactic and semantic features of a truth bearer are not always sufficient to determine whether it is paradoxical and the fact that paradoxicality affects truth value (on some views); see Scharp (RB) for discussion.

[43] See Scharp (FTT) for a defense of this criticism.

[44] Ibid.

[45] See Scharp (TI) for a defense of this criticism.

we assume that natural language truth predicates are LS truth predicates, one can still 'import' a sentence that is potentially paradoxical for a theory of truth into revised natural language to which the theory applies—it just takes a bit more work. Assume that T is a version of Kripke's theory of truth that is intended to explain truth-in-English* and that T is restricted to English*.[46] English* is like English except that it does not contain any of the resources that can be used to construct revenge paradoxes for T. Let L be a language that does contain resources that could be used to construct revenge paradoxes for T (i.e. there are sentences of L that are potentially paradoxical for T—if T had applied to L, then these sentences would be paradoxical for T). Let ϕ be one of these sentences. The direct way of 'importing' a potentially paradoxical sentence into English* no longer works. That is, 'ϕ is true-in-English*' is not paradoxical for T because it is false-in-English*. However, consider another language, M such that: (i) M has a sentence, ψ that is a translation of ϕ, (ii) ψ has certain specifiable empirical features (e.g. it is the only sentence written on a certain blackboard), and (iii) M has a sentence θ that attributes truth-in-M to ψ by appeal to its empirical features (e.g. 'the sentence on the blackboard is true-in-M). Assume that: (i) English* has a sentence σ that is a translation of θ, and (ii) that English* has a sentence ζ that attributes truth-in-English* to σ.[47] Although it might not seem like it, if T applies to ζ, then ζ is paradoxical for T. Here is the argument. If ϕ is paradoxical for a theory of truth-in-L that applies to ϕ, then θ is paradoxical for a theory of truth-in-M that applies to θ. If θ is paradoxical for a theory of truth-in-M that applies to θ, then ζ is paradoxical for a theory of truth-in-English* that applies to ζ.[48] The argument relies on a combination of inter-linguistic truth attributions and empirically paradoxical sentences to show that it is impossible to eliminate from a natural language all the sentences that are potentially paradoxical for a given theory of truth (as long as the natural language has the features we take all natural languages to have).

Objection 8: Although the inconsistency problem is a genuine problem for a theory of truth, the self-refutation problem does not really pose a threat. In particular, it is acceptable for a theory of truth to imply that some of its consequences are gappy. Of course, the theory will end up being gappy as well, but that is acceptable too. In order for this sort of approach to be workable, one must accept that gappy

[46] This restriction is not trivial because languages other than English* contain truth-in-English* predicates (e.g. German has 'wahr-auf-Englisch*'). Assume that sentences of languages other than English* are false-in-English*.

[47] All English* needs in order to have a translation of θ is a truth-in-M predicate and a way of specifying the empirical features of ψ, which it must have if it is to be a workable natural language. If English* does not have a truth-in-M predicate, then all English* truth attributions to sentences of M that express truth-in-M are false-in-English—an intolerable result for a descriptive theory.

[48] See Scharp (TI) for a more detailed version of this argument.

sentences can be assertible, but that can be accommodated as well. Given that the other options are either a restricted (hence, unacceptable) theory or an inconsistent (hence, unacceptable) theory, a gappy theory doesn't seem so bad.[49]

Reply 8: There are several problems with this sort of theory. The most significant problem is that we are no longer able to use truth in our assessment of a theory. When one presents a theory, one wants one's theory to be correct—to get it right. That is, one wants one's theory to be true. Presumably, a theorist who offers the type of theory in question has the same desire. If asked, 'is your theory correct?,' he would probably say, 'yes'. However, given that the theory is gappy by its own lights, the theorist is unable to make the move from 'I have it right' to 'my theory is true'. We use truth as a measure of theoretical correctness and a theory of the type in question does not respect this aspect of our use of truth.

The second problem is that the approach to the liar paradox that involves accepting a theory of truth that implies that it is gappy depends on either restricting the theory so that it does not apply to certain languages or denying that certain linguistic resources are meaningful (i.e. an unintelligibility maneuver). Recall that a theory for which the self-refutation problem occurs implies that (1) (a liar sentence) is gappy, that (2) (a revenge liar sentence) is gappy, and that '(2) is true if and only if (2) is either false or gappy' (a consequence of the theory) is gappy.

The problem occurs with certain sentential operators that force one into a two-valued scheme. Recall the sentence:

(5) (5) is Xnot true.

Just as in the case of a liar sentence, if we assume that (5) is true or that (5) is false, then we get a contradiction. If we assume that (5) is gappy, then we can derive that (5) is false. Thus, an approach of this sort cannot handle sentences like (5). The only options are to restrict the theory so that it does not apply to sentences like (5) or claim that the linguistic resources required to construct sentences like (5) (e.g. exclusion negation) are meaningless. As I have argued, neither option is remotely plausible.

13.4 Inconsistency Arguments

In this section, I present three arguments for an approach to truth on which truth is an inconsistent concept. They are presented in the order of increasing specificity for their conclusions. The first argument is an abductive argument: an inconsistency

[49] See Maudlin (2004) for an example of a theory of this type.

theory of truth provides the best explanation of all the paradoxes associated with truth (including all the various revenge paradoxes). This argument supports any theory of truth that implies that truth is an inconsistent concept. The second argument is based on the revenge argument given in the last section. It supports most inconsistency theories of truth, but not all. The third argument appeals to the claim that a particular strategy for explaining our inconsistent concept of truth avoids all the revenge paradoxes and, hence, allows for a theory of truth that does not have to be restricted. This argument supports only inconsistency theories of truth that meet very specific conditions. In the following section, I present an overview of one such theory.

Before presenting those arguments, I want to discuss concepts and what I mean when I say that truth is an inconsistent concept. Roughly, I think of concept application on the lines of belief formation or assertion. For example, I apply the concept **scab**[50] to some object α if I am prepared to assert 'α is a scab' or I believe that α is a scab. An important part of my account of concepts is that there are rules that govern the employment of a concept. A person who possesses a certain concept and is committed to employing it is committed to following the rules for the employment of that concept. One such rule is that the concept **scab** should be applied to scabs and it should not be applied to things that are not scabs.[51] One can think of these rules as *constitutive* in the sense that if a person utters a word that expresses the concept **scab**, then that person is committed to following the rules for the employment of **scab**. By 'committed to following the rules' I do not mean that the person actually acts in accordance with these rules or that he explicitly endorses them; rather, I mean that the person *ought* to follow them—he is obligated to follow them—whether he explicitly endorses them or not (of course, it is not the case that everyone actually obeys the constitutive principles for a given concept—some disobey out of ignorance, others do so on purpose).[52]

An *inconsistent concept* is one whose constitutive rules are incompatible in the sense that they dictate that the concept both applies and disapplies to some entities.[53] The rules

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

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

[52] My commitment to constitutive rules for concepts places me in the tradition of meaning-constitutive accounts of concepts. However, not all the members of this tradition agree on rules as the relevant constitutive element. Some constitutive accounts choose the possession of propositional attitudes, the truth of theories, the validity of implications, etc. See Peacocke (1992) for an example; however, Peacocke treats a concept's constitutive principles as those possessors of the concept are committed to following, whereas I treat them as those employers of a concept are committed to following. One can possess a concept without employing it; see below for the role this claim plays in my account.

[53] I have yet to find a good antonym for 'applies'.

for the employment of an inconsistent concept impose conflicting commitments on the employers of that concept. Thus, the employer of an inconsistent concept cannot follow the rules for the application of that concept in all circumstances.[54] Consider an example:

(7a) 'rable' applies to x if x is a table.
(7b) 'rable' disapplies to x if x is red.[55]

Rable is an inconsistent concept. Someone who possesses **rable** might run into difficulty employing it because it both applies and disapplies to red tables. When confronted with a red table, an employer of **rable** will be unable to satisfy the demands it places on her. It might seem that someone could employ **rable** without trouble as long as she avoids red tables. However, even if an employer of **rable** never encounters a red table, the concept still poses a problem for her because inconsistent concepts pose a normative problem for their employers. Someone who chooses to employ **rable** *should* apply it to tables and *should* disapply it to red things. These are conceptual norms to which the employer has decided to bind herself. Thus, an employer of **rable** has committed herself to obeying incompatible rules even if she never encounters a red table.

Inconsistency theories of truth imply that truth is an inconsistent concept. That is, they imply that truth has certain constitutive rules that govern its employment and these rules are incompatible in the sense that they dictate that truth both applies and disapplies to certain items. I am not concerned with arguing about which rules are constitutive for truth, but it seems to me that the truth rules are constitutive of it along with the principle of mono-aletheism (i.e. no truth bearer is both true and false simultaneously). Liar sentences are among the items to which truth both applies and disapplies by virtue of the fact that the rules governing its employment are incompatible.

These sketchy remarks about truth as an inconsistent concept are bound to raise more questions than they answer, but they will have to do for now. A complete theory of inconsistent concepts including a logic, a semantic theory, and a pragmatic theory is beyond the scope of this Chapter. Likewise, the details of how such a theory should be applied to the case of truth will have to wait for another occasion. (However, I briefly mention some of my views on both matters in section 4.)

The first argument is that if one decides to treat truth as an inconsistent concept, then one has available a satisfying explanation of the current situation in truth studies. That is, one can explain why other theories of truth face revenge paradoxes, both

[54] See Chihara (1979, 1984), Yablo (1993a), and Eklund (2002) for similar views on inconsistent concepts.

[55] I say that a concept *applies* to the members of its extension and *disapplies* to the members of its anti-extension.

inconsistency problems and self-refutation problems. No other theory of truth has managed to do this.[56]

The explanation for why theories of truth that imply truth is a consistent concept face revenge paradoxes or self-refutation problems is straightforward. Our concept of truth is inconsistent in the sense that its constitutive principles, the truth rules, are incompatible. That is, there are objects that these rules classify as both true and not true. We can call the set of such objects the *overdetermination set* for truth. All the paradoxical sentences considered so far are members of the overdetermination set for truth.[57] Any theory of truth that implies that truth is a consistent concept and that includes these principles is inconsistent and can be rendered consistent only by restricting it. If a theory of truth implies that some of the sentences in the overdetermined set for truth are gaps, then its fate depends on which of these sentences it classifies as gaps. Recall that many of the members of the overdetermination set for truth are truth attributions, and no matter what truth status (e.g. true, false, gappy, etc.) one assigns them, they are consequences of the assignment. No matter whether one's theory of truth classifies these paradoxical sentences as true, false, or gappy, some of these paradoxical sentences are consequences of the theory. Thus, if a theory of truth implies that all the sentences in the overdetermined set for truth are gaps, then the theory implies that some of its consequences are gaps. On the other hand, if a theory of truth does not classify some of these sentences as gaps, then the truth rules imply that they are both true and not true. On the first option, the theory is self-refuting, while on the second, it faces an inconsistency problem. Therefore, both types of revenge paradoxes can be explained if we assume that truth is an inconsistent concept.

In section 2, I argued that theories of truth that validate the truth rules face revenge paradoxes. If we admit that truth is an inconsistent concept, then we can explain why this occurs. Therefore, by accepting that truth is an inconsistent concept, we arrive at a deeper explanation for why theories of truth that validate the truth rules fail are unacceptable.

The following is a summary of the first argument for treating truth as an inconsistent concept:

(i) If we assume that truth is an inconsistent concept, then we can explain the presence of the liar paradox and the presence of revenge paradoxes.

[56] See Field (forthcoming) and Glanzberg (2005) for the only alternative explanations of which I am aware; I consider them in a reply to an objection below.

[57] It seems to me that truth-tellers (e.g. sentence τ, 'τ is true', is a truth-teller) are in the *under*determination set for truth, but these sentences are not paradoxical and none of my claims or arguments hang on this opinion.

(ii) The inconsistency explanation of the liar paradox and the revenge paradoxes is better than any of the others.

∴ (iii) Probably, truth is an inconsistent concept.

Only by admitting that truth is an inconsistent concept can we satisfactorily explain the most significant feature of our long battle with the liar paradox.

The second argument for the claim that truth is an inconsistent concept depends on the revenge argument from the previous section: if a theory of truth implies that the truth rules are valid, then it is inconsistent, self-refuting, or restricted. Given that any theory of truth that is inconsistent, self-refuting, or restricted is unacceptable, any theory of truth that implies that the truth rules are valid is unacceptable. That is the conclusion of the revenge argument. I have assumed that we should accept that the truth rules are constitutive of truth—that they govern the way truth should be employed. Again, this claim counts as an assumption in this chapter because I do not have the space to argue for it here; however, there are good reasons to accept it. Thus, on my view: (i) if a theory of truth is acceptable, then it implies that the truth rules are constitutive for truth; and (ii) if a theory of truth validates the truth rules, then it is unacceptable. There seems to be very little wiggle room here. However, there is an additional assumption that connects the two conditionals: if a theory of truth implies that the truth rules are constitutive of truth, then it implies that they are valid. I reject this claim. I accept that if a concept is not defective (i.e. its constitutive principles are compatible with one another and empirical facts), then its constitutive principles are valid (or true). However, for inconsistent concepts, it is not so straightforward. One of the key aspects of the account of inconsistent concepts I endorse is that one or more of the constitutive rules governing the concept are invalid.[58] Recall that the constitutive principles for a concept are those that one commits oneself to following if one is committed to employing the concept. However, inconsistent concepts should not be employed. Of course, they are still possessed, but there are plenty of concepts we possess even though we do not employ them. Only if one is committed to employing a concept does one assume that its constitutive inference rules are valid. Thus, I accept that the truth rules are constitutive of truth, but it is not the case that they are all valid. This result allows anyone with a suitable account of inconsistent concepts to avoid the revenge argument. The following is a summary of the second argument for treating truth as an inconsistent concept:

(i) If a theory of truth is acceptable, then it implies that truth obeys the truth rules.

[58] This claim is a result of the fact that I explain inconsistent concepts in terms of confused concepts. Thus, I claim that one should explain an inconsistent concept by appeal to a set of component concepts. Each component concept obeys only some of the constitutive principles for the inconsistent concept in question. An inference rule for the inconsistent concept is valid if and only if it is valid for each of the component concepts.

(ii) If a theory of truth implies that truth is a consistent concept and it implies that truth obeys the truth rules, then it implies that the truth rules are valid.

(iii) If a theory of truth implies that the truth rules are valid, then it is inconsistent, self-refuting, or restricted.

(iv) If a theory of truth is inconsistent, self-refuting or restricted, then it is unacceptable.

∴ (v) If a theory of truth is acceptable, then it does not imply that truth is a consistent concept.

Only an inconsistency theory of truth on which inference rules that are constitutive for a concept need not be valid has a chance of being an acceptable theory of truth.

The third argument justifies only a small number of inconsistency theories of truth. In short, the argument is: if one assumes that truth is an inconsistent concept, then one can construct a theory of truth that does not face any revenge paradoxes and so does not have to be restricted in any way. Given that there are no other theories of truth that are really revenge-free, the fact that an inconsistency theory of truth can accomplish this feat is a strong reason to accept it. If I am right that theories of truth that are restricted to avoid revenge paradoxes are unacceptable and that no other purportedly acceptable theories of truth avoid the revenge paradoxes, then the theory I endorse is the only acceptable theory of truth to have ever been proposed. It should not come as a shock that I cannot provide a convincing argument for the claim that the theory I endorse is genuinely revenge-free—for I do not even have the space to present the theory in detail. Moreover, I am unable to fully justify my claim that all the other purportedly acceptable theories of truth do have to be restricted in some way. Nevertheless, the discussion in the preceding section should motivate (to some degree) the claim about the other theories, while the discussion in the following section should motivate (to some degree) the claim about my own theory.

The following is a summary of the third argument for treating truth as an inconsistent concept:

(i) If a theory of truth is acceptable, then it is not restricted.

(ii) A particular inconsistency theory of truth T does not face revenge paradoxes and so is not restricted.

(iii) Of the proposed theories of truth, those other that T are either antecedently implausible or restricted.

∴ (iv) Of the proposed theories of truth, only T is acceptable.

Only the theory of truth I propose (or one that is relevantly similar to it) is acceptable.

Objection 1: There are no inconsistent concepts. Any attempt to introduce a term that behaves according to incompatible rules fails to introduce a meaningful term

at all. Thus, it is impossible that a term obeys incompatible rules of employment. One reason for thinking this is that interpretation requires one to use the logic one endorses when interpreting another. Thus, it is inappropriate to ever attribute an inconsistent concept to someone, since the interpreter would have to attribute something that defies the logic she endorses.[59] Moreover, even if one could introduce a term that obeys incompatible rules, it would be overdetermined for every item, so it would be unemployable.[60]

Reply 1: First, the claim that we interpret others as if they endorse our logical standards is simply false. If it were true then there would be no distinction between criticizing someone for failing to follow an inference rule she endorses and criticizing someone for endorsing the wrong inference rule. It is obvious that there is such a distinction and it plays an important role in philosophical discussions. Second, charity can cut both ways. One might simply introduce an inconsistent concept, begin using it, and describe it as inconsistent (I did this with the concept **rable**). It seems to me that it would be quite difficult to go on interpreting someone who does this as if they had misunderstood their own stipulative definition and their claims about it. Indeed, one might give an account of all the relevant factors in charitable interpretation and present two situations, one in which the weighted sum of all the factors is higher than that of the second, while in the first one attributes an inconsistent concept, but in the second one does not. The point here is that attributing an inconsistent concept is sometimes the most charitable thing to do. No matter what constraints one imposes on charitable interpretation (except, of course, a conceptual consistency constraint), there will be situations in which it is more charitable to attribute an inconsistent concept.

I agree that a major problem for a theory of inconsistent concepts is showing that a concept can be both inconsistent and employable (i.e. not overdetermined for every item). The theory I endorse accomplishes this, in part, by employing a relevance logic to evaluate arguments with sentences containing truth predicates; however, a complete account is beyond the scope of this chapter.[61]

Objection 2: The account of inconsistent concepts depends on the claim that each concept has some constitutive principles. However, there are good reasons to reject this claim, which include: (i) it commits one to analytic truths;[62] and (ii) it implies that there is a well-defined distinction between changes in meaning and changes in belief.[63]

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

[60] See Gupta and Belnap (1993: 13–15) for this objection; see also Chihara (1984) for discussion.

[61] Note that one needs a two-part solution to the problem of overdetermination—one needs some sort of paraconsistent logic on which it is not the case that everything follows from a contradiction, and one needs a logic on which it is not the case that all the constitutive rules for the concept in question are valid.

[62] See Fodor and Lepore (1994).

[63] See Davidson (1974).

Reply 2: I agree that many meaning-constitutive accounts of concepts face those problems, but mine does not. I do not claim that the meaning-constitutive principles concern concept *possession*; rather, they concern concept *employment*. One can possess a concept without being committed to employing it. Indeed, I suggest that although we possess the concept of truth, we should not employ it. On the account I offer, simply possessing a concept does not commit one to its meaning-constitutive principles at all; only if one is committed to employing a concept is one committed to its meaning-constitutive principles. Without this distinction, it seems impossible to give an adequate account of inconsistent concepts.

Although other views on meaning-constitutive principles do imply that these principles are analytic (i.e. true or valid by virtue of their meaning alone), the one I offer does not. Indeed, it is my view that *any* sentence that expresses an inconsistent concept has no truth-value (because I think that truth itself is inconsistent, when I use 'truth value' I am appealing to the concepts of truth I offer as replacements for our inconsistent concept).[64] Thus, no meaning-constitutive principles are analytic and some are not even true or valid.[65] Moreover, one can endorse a meaning-constitutive account of concepts and admit that there is no principled way of distinguishing between meaning change and belief change. The reason usually given for this criticism is that if one's word means X, then one has to believe the associated meaning-constitutive principle. However, the account I offer is externalist in a certain sense; if one employs a certain concept, then one is committed to its constitutive principles, but one need not have the associated beliefs. That is, we can imagine a situation in which a person employs the concept *elm* and the concept *beech* and she is committed to the respective constitutive principles governing those concepts, but she doesn't have any beliefs that distinguish elms from beeches. Indeed, one can employ a certain concept even though one denies its constitutive principles (e.g. theorists who offer approaches to the liar paradox on which the truth rules are not constitutive).

Objection 3: There are other inconsistency theories of truth. Why aren't they at least as good as the one I offer?

Reply 3: There are several inconsistency theories of truth on the market and I cannot do justice to all of them here; I confine my comments to pointing out differences

[64] Otherwise, one would not be able to distinguish between concept possession and concept employment. For example, if some sentence p that expresses an inconsistent concept has a truth-value, then one should be able to assert that it has that truth-value. If one can assert that p has a certain truth-value, then (because of the truth rules) one is committed to either p or its negation. Thus, if p has a truth-value, then it is impossible to avoid employing the inconsistent concept it expresses.

[65] One might worry that if a concept is not defective, then its meaning-constitutive principles are analytic. On the contrary, in the case of a non-defective concept, a meaning-constitutive principle is true or valid, but it is true or valid by virtue of both its meaning and the fact that the concept is not defective (which depends on empirical facts).

between my theory and the rest and explaining why I prefer my account. None of what follows should be treated as real criticism. Yablo has a theory of inconsistent concepts that is based on an account of circularly defined concepts.[66] Eklund has a theory of inconsistent concepts as well, but he focuses on entire languages (I take it that an inconsistent language is just a language that has a term that expresses an inconsistent concept). He suggests that by virtue of our semantic competence, we accept the constitutive principles of an inconsistent concept and he offers a semantics for inconsistent languages on which an acceptable assignment of semantic values to expressions of a language, L, is one that makes true a weighted majority of the constitutive principles for L.[67] The biggest problem I have with the theories Yablo and Eklund propose is that they require that we continue to employ our inconsistent concept of truth. That is, they imply that it is acceptable to employ an inconsistent concept even after one has discovered that it is inconsistent. Indeed, both Yablo's theory and Eklund's theory appeal to the concept of truth. Because their theories appeal to the inconsistent concept of truth, anyone who accepts either theory has to employ an inconsistent concept. I think that is unacceptable. It is my view that inconsistent concepts should be replaced with consistent ones; they aren't fit for employment. The rationale for this view is simple: other things considered, one should avoid undertaking incompatible commitments. Thus, if one discovers that a concept is inconsistent, one should stop employing it if one can.

In addition, the fact that these theories appeal to the inconsistent concept of truth causes problems for them that are similar to the revenge paradoxes. For example, if a philosopher accepts Eklund's theory of truth, then she also commits herself to using truth according to principles that are incompatible. Thus, she commits herself to both applying and disapplying truth to certain sentences. That is similar to the inconsistency problem. Moreover, Eklund's theory still has to classify paradoxical sentences and some of these will be consequences of the theory. If it classifies them as false, then his theory has false consequences. If it classifies them as gaps, his theory has gappy consequences. Either way, his theory is self-refuting. Similar remarks hold for Yablo's theory.

Another theory that might seem like an option is dialetheism—the view that some contradictions (e.g. '(1) is true and (1) is false') are true. Given what I have said about inconsistency theories of truth, dialetheism need not count as one. That is, the dialetheist can just deny that mono-aletheism (i.e. no truth bearer is true and false simultaneously) is a constitutive principle for truth. According to this version of dialetheism, the rules governing our concept of truth are compatible (as far as I understand it, all contemporary dialetheists fall into this category); they just imply that truth

[66] Yablo (1993a, 1993b).

[67] Eklund (2002, forthcoming). See also Azzouni (2003) and Patterson (2006) for similar theories; I can't give them all the space they deserve.

and falsity overlap in some cases.[68] However, one might use dialetheism as a part of a theory of inconsistent concepts. On this version of dialetheism, mono-aletheism is a constitutive principle for truth; hence, the constitutive principles for truth are incompatible. As a theory of inconsistent concepts, dialetheism implies that some sentences expressing inconsistent concepts are both true and false (e.g. 'the red table is rable'). It is unclear to me whether the two versions of dialetheism (i.e. truth is a consistent concept that doesn't obey mono-aletheism vs. truth is an inconsistent concept that does obey mono-aletheism) are different in a substantive way. Instead of worrying about this issue, let me say that I reject the first version of dialetheism because it does not respect one of the constitutive principles for truth (as I said in section 2, this isn't an objection by itself, but it does point the way toward one). Although the second version of dialetheism (as a theory of inconsistent concepts), avoids this problem, it still requires one to accept that some truth-bearers are both true and false (which seems like a consequence we should avoid if we can). Moreover, this version of dialetheism has the same problem as other theories of inconsistent concepts: because it appeals to the inconsistent concept of truth, anyone who accepts the theory has to employ an inconsistent concept. It also cannot deal with certain non-monotonic sentential operators, like exclusion negation.

Objection 4: There are at least two well-developed explanations of the revenge paradox phenomenon, one from Hartry Field and one from Michael Glanzberg. Why is the explanation I offer superior to the ones they offer?

Reply 4: Let me begin the reply by saying that Glanzberg and Field each have interesting, complex, and subtle approaches to the liar paradox and there is no way I can do justice to either of them in this reply. One should not take what I write here to be something like definitive criticisms; instead, I aim to point out relevant differences between our respective explanations.

I begin with Glanzberg, who offers a context-dependence approach to the liar. However, instead of claiming that truth predicates are explicitly context dependent, Glanzberg argues that sentences that contain truth predicates display an implicit context dependence that is due to the presence of quantification. Glanzberg offers a theory of background domains of propositions for the quantifiers involved, which includes an infinite hierarchy of domains and no 'biggest' domain.[69]

If one accepts Glanzberg's theory, then one has to admit that there is no unrestricted quantification. Indeed, one has to accept that we can express the notion of truth-in-a-context and we can even quantify over contexts to a limited degree, but we cannot express an unrestricted notion of truth. 'One way or another, hierarchical theories all require that speakers cannot in any one instance express the entirety of a unified concept of truth,' (Glanzberg 2004: 289). He argues that the sort of fragmentation we

[68] See Priest (1987, 2006), and the papers in Priest, Beall, and Armour-Garb (2004).

[69] Glanzberg (2001, 2004, 2005).

see in our concept of truth is familiar to us (i.e. it occurs in the concept of mathematical proof as well) and that it occurs because truth fails to be closed under reflection.

Glanzberg's defense of this feature is based on the idea that any characterization of truth permits one to reflect on the truth of the characterization, and this reflection both shows that the initial characterization is inadequate and points the way toward a stronger one. This process of reflection is unending; hence the infinite hierarchy of contexts.[70] The motivation for this view comes from what has been called the strong liar reasoning. Consider the following sentence:

(8) (8) is not true.

The partiality approach to the liar implies that (8) is gappy. We know that if a sentence is gappy, then it is not true. Thus, the partiality approach implies that (8) is not true. Hence, the partiality approach implies that '(8) is not true' is true; therefore, it implies that (8) is true.[71] It is by reflection on the way the approach classifies (8) that drives us to conclude that (8) is true after all. The claim, ⟨if the partiality approach implies that ⟨p⟩ is true, then ⟨p⟩ is true⟩, is similar to what is called a reflection principle. It states something about a formal theory that cannot be captured by the formal theory on pain of contradiction.[72]

Glanzberg argues that one can begin with a basic formal theory of truth, formulate a reflection principle for that theory, which illustrates the theory's inadequacy, and arrive at a new formalization of the theory that effectively incorporates the reflection principle. We can continue this process to arrive at a transfinite hierarchy of formal theories of truth, which is analogous to the hierarchy of contexts for truth attributions. He claims that truth is a *Kreiselian concept* in this sense: any formal theory of truth points the way to a stronger formal theory, and the process of theory construction is unending.[73]

Glanzberg's point is that what seem to be revenge paradoxes are really just the effects of the Kreiselian aspect of truth. A theory of truth should not be expected to treat as true the claim that its consequences are true. Nor should a theory of truth be found lacking if the result of conjoining a reflection principle to it results in an

[70] This view about the relation between reflection and revenge seems to stem from some of Kripke's remarks: 'Such semantical notions as "grounded", "paradoxical," etc. belong to the metalanguage. This situation seems to me to be intuitively acceptable; in contrast to the notion of truth, none of these notions is to be found in natural language in its pristine purity, before philosophers reflect on its semantics (in particular, the semantic paradoxes). If we give up the goal of a universal language, models of the type presented in this paper are plausible as models of natural language at a stage before we reflect on the generation process associated with the concept of truth, the stage which continues in the daily life of nonphilosophical speakers' (Kripke 1975: 714).

[71] See Burge (1979) for discussion of the strong liar reasoning.

[72] See Feferman (1991) for an overview of reflection principles.

[73] Glanzberg (2005).

inconsistent theory. These phenomena are just consequences of the fact that truth is a Kreiselian concept.

There are several places at which I disagree with Glanzberg's analysis. The first is that I do not find the strong liar reasoning compelling. Because strong truth does not have truth-value gaps, and the strong liar reasoning involves a move from ⟨⟨p⟩ is gappy⟩ to ⟨⟨p⟩ is not true⟩, (8) should be read as: (8) is Xnot weak true. We already know that partiality approaches have troubling handling sentences like this, but the trouble has nothing to do with reflection on how the theory of truth in question classifies (8). In fact, most partiality approaches to the liar are based on fixed-point constructions and so have no consequences for sentences like (8) at all. Thus, there is no reason to think that one derives a contradiction only by assuming that the theory implies that (8) is gappy. Therefore, the view that reflection on the dictates of a theory of truth has any special role to play in reasoning about the status of paradoxical sentences seems to be a mistake.

A second issue is that there seems to be no reason to think that the concepts described by each of the formal theories Glanzberg identifies have anything in common. In particular, I see no reason to think that they are all 'more or less' theories of truth. However, Glanzberg claims that each of the formal theories provides a rough characterization of the unified concept of truth. The problem with this view is that on Glanzberg's own account, it is impossible to express the unrestricted notion of truth each of these theories is supposed to describe. Thus, we have this concept of truth, but we can never actually use it. That sounds fairly counterintuitive. Moreover, if Glanzberg is right, then it is impossible to arrive at a theory of truth that correctly and completely describes our concept of truth. The best we can achieve is stronger and stronger theories that are always lacking.

On the other hand, Field presents a sophisticated version of the partiality approach that is designed to handle not just liar sentences but certain revenge sentences as well. Field shows how to introduce a conditional into a three-valued scheme that obeys most of the intuitive principles we take to govern conditionals and acts just like a material conditional in classical contexts (i.e. those for which excluded middle is assumed). It is hard to overestimate the significance of this accomplishment; for the first time, we have a three-valued logic with a real conditional—a three-valued logic in which we can reason. Field then defines a determinacy operator in terms of the conditional. He claims that liar sentences are indeterminate (not determinately true and not determinately false). Given what I have said already about revenge paradoxes, it should be obvious that a sentence like the following poses a problem for an approach like Field's:

(6) (6) is either false or indeterminate.

Field's determinacy operator iterates non-trivially, so he can say that (6) is not determinately determinately true and not determinately false (determinate determinate

truth is weaker than determinate truth, but not so for determinate falsity). We can say that a liar sentence is indeterminate$_0$ (i.e. it is not determinately true and not determinately false) and that (6) is indeterminate$_1$ (i.e. it is not determinately determinately true and not determinately false). Field shows how to construct a hierarchy of determinacy operators (by iterations of the primitive one defined in terms of the conditional), which are used to define a hierarchy of determinate truth predicates and a hierarchy of indeterminateness predicates. As a final touch, he proves that his theory of truth is expressible in the artificial language he constructs (which has the expressive resources to handle set theory) without giving rise to revenge paradoxes. Thus, his theory does not appeal to a distinction between object language and metalanguage; it can be applied to the very language in which it is formulated. Most impressive.[74]

In section 2 on the revenge argument, I argued that partiality approaches to the liar paradox face revenge paradoxes. How does Field avoid that argument? He rejects the excluded middle for all of the indeterminateness predicates he defines. That is, he rejects '(6) is indeterminate$_0$ or (6) is not indeterminate$_0$'. Thus, all of his indeterminacy predicates are analogous to truth (on his view) in that they are partially defined. Of course, as one goes further and further up the hierarchy of indeterminacy predicates, one gets closer and closer to a completely defined indeterminateness predicate, but Field argues convincingly that one never gets there. That is, it is impossible to define a completely defined indeterminacy predicate in the language he constructs. Thus, he never has to deal with a real revenge paradox, like:

(6*) (6*) is either false or indeterminate*,

where 'indetermiante*' is completely defined (i.e. every sentence is either indeterminate* or not indeterminate*). Field's theory clearly cannot handle sentences like (6*). What does he say about them?

Field argues (convincingly in my view) that one need not use any such linguistic expression to formulate his theory (his semantic theory uses a completely defined notion of semantic value, but it is relative to a model and so cannot be used to construct a revenge paradox). In addition, he argues against the claim that the artificial language he constructs avoids revenge paradoxes only because it has expressive limitations. Instead, he claims that sentences like (6*) are unintelligible. Indeed, Field argues, any linguistic expression that is not in his artificial language and seems to give rise to a revenge paradox (e.g. 'indeterminate*') is unintelligible; however, by 'unintelligible' he does not mean *meaningless*:

I don't want to deny that we have these notions; but not every notion we have is ultimately intelligible when examined closely. A large part of the response to the counterintuitiveness

[74] See Field (2003a, 2003b, 2005a, 2005b, forthcoming).

qualm will be an argument, in Part Four, that the notion of 'the' hierarchy of iterations of D has a kind of inherent vagueness that casts doubt on there being a well-behaved notion of 'Dα-true for every α'; and without that there is no reason to suppose that there is a well-behaved notion of 'determinately true in every reasonable sense of that term'. The apparent clarity of such notions is an illusion. (Field, forthcoming c: §11).

Field then argues that there is no way to extrapolate from the hierarchy of determinateness predicates to define a well-behaved (i.e. intelligible) notion of hyper-determinateness.[75]

I am willing to admit that if we have only the resources provided by Field's artificial language, then we will be unable to define a well-behaved notion of hyper-determinateness. However, it seems to me that this point does little to quell the revenge worries. The problem is *not*: how can we use the resources Field gives us to generate a paradox his theory cannot handle? The problem is: we have a notion of determinateness that obeys excluded middle (i.e. what Field calls hyper-determinateness) and one cannot express this notion in Field's artificial language. Thus, Field avoids revenge only by an expressive limitation on his language.

I assume that it is obvious how Field would respond. He would probably claim that his artificial language can express any *intelligible* notion we have. Furthermore, he might continue, the problem of revenge paradoxes that I keep pressing is a problem that arises only when truth, which is intelligible, is combined with other resources (e.g. exclusion negation, other non-monotonic sentential operators, hyper-determinateness operators, etc.), which are not intelligible. Thus, it is not that truth is responsible for these revenge paradoxes; rather, truth has been keeping company with some troublemakers who are responsible. I address this reply in the next objection.

Before doing so, a summary is in order. I have argued that there are two kinds of revenge paradoxes: self-refutation problems; and inconsistency problems. Glanzberg addresses self-refutation problems, which confront theories of truth that imply that they are Xnot true. He argues that this kind of revenge paradox has its source in the fact that truth is a Kreiselian concept (i.e. it is not closed under reflection). However, Glanzberg does not explain or even address inconsistency problems, and there are good reasons to doubt his explanation of the self-refutation problem. On the other hand, Field addresses inconsistency problems and argues that these sorts of revenge paradoxes arise when what are ultimately unintelligible—read that as inconsistent or not well-defined—concepts (e.g. hyper-determinateness) are combined with truth. However, Field does not explain or even address self-refutation problems, and there are good reasons to doubt his explanation of the inconsistency problem. In contrast

[75] Field also discusses what he calls 'model-theoretic revenge', but it is distinct from the sort of worry that I've pressed in this chapter; see Field (forthcoming: §9).

to both Glanzberg and Field, I offer an explanation of both types of revenge paradoxes, and my explanation of each type is superior to the one offered by Glanzberg and by Field, respectively.

Objection 5: As Field suggests, we should assume that the other items involved in the revenge paradoxes are defective instead of assuming that truth is defective. Given the importance and centrality of truth and the relative unimportance of these other items, we should prefer a theory on which truth is an acceptable concept and the others are defective.

Reply 5: There are several issues to consider when deciding which linguistic items should be blamed for the paradox. One issue involves the sort of explanation we get. The objector suggests that we should blame exclusion negation, and blame all the other non-monotonic sentential operators, and blame the conditional, and blame completely defined gaphood predicates, and blame idempotent determinacy operators, and blame quantification over partially defined gaphood predicates, and blame paradoxicality predicates, and blame groundedness predicates, and blame truth expressions that are not language-specific . . . the list goes on. I suggest that we should blame truth. That's it. Thus, my explanation is much simpler. It is also much more plausible. We can construct artificial languages that contain the outlaw linguistic expressions and they are perfectly well-behaved as long as they do not contain truth predicates (or related semantic terms). Of course, we can also construct artificial languages with truth predicates that are perfectly well behaved as long as they do not contain the outlaw linguistic expressions. However, the difference is that there are many different ways to construct revenge paradoxes; one involves truth and exclusion negation, one involves truth and another non-monotonic sentential operator, one involves truth and the conditional, one involves truth and an idempotent determinacy operator, etc. Exclusion negation is not involved in each case, nor are any of the other outlaw linguistic expressions. However, truth is involved every time. Truth is the only suspect that has no alibi—it is present at every crime scene; none of the others is. It does not take a Holmes, or a Spade, or a Columbo to identify the perpetrator; even a Wiggum could get this one right.

However, there is another, even more compelling reason to prefer my approach. As I said in reply to objections 6 and 7 of section 2, as long as one has a truth predicate that obeys the truth rules (even a language-specific one) and some minimal resources (e.g. common descriptions), one can 'import' a revenge paradox into the language by way of inter-linguistic truth attributions. Field avoids this problem only because he does not consider other languages at all. Thus, even if we follow Field's advice (i.e. blame the outlaw linguistic resources for the liar and revenge paradoxes and revise our language so that it does not contain any outlaw linguistic resources), then we still have to restrict the theory so that it does not apply to certain sentences of the revised language. Thus, Field's strategy does not really secure a language and a

theory such that the theory both applies to and is expressible in the language. Hence, the 'blame everything but truth' strategy does not work. Not only is my strategy simpler and more plausible, it is the only one that works.

13.5 The Way Forward

Inconsistent concepts are rarely a topic of contemporary philosophical discussions, and there are, to my knowledge, no detailed theories of inconsistent concepts.[76] Given that the central claim of the account of truth I offer is that truth is an inconsistent concept, I am badly in need of an adequate theory of inconsistent concepts. My strategy is to first construct such a theory and then apply it to the case of truth.

A central claim of the account of inconsistent concepts I offer is that inconsistent concepts are confused. A concept is confused if and only if anyone who employs it fails to properly distinguish between two or more entities. A classic example is the concept of mass as it was employed in Newtonian mechanics. In Newtonian mechanics, physical objects have a single physical quantity: mass. According to this theory, the concept of mass obeys the two laws (which are considered equally fundamental): (i) mass = momentum/velocity; and (ii) the mass of an object is the same in all reference frames. However, in relativistic mechanics, physical objects have two different 'kinds' of mass: proper mass and relativistic mass. An object's proper mass is its total energy divided by the square of the speed of light; an object's *relativistic mass* is its non-kinetic energy divided by the square of the speed of light. Although relativistic mass = momentum/velocity, the relativistic mass of an object is not the same in all reference frames. Contrariwise, proper mass ≠ momentum/velocity, but the proper mass of an object is the same in all reference frames. Thus, relativistic mass obeys one of the laws for mass, and proper mass obeys the other. A person who employs the concept of mass thinks that there is one thing (mass), when there are really two (relativistic mass and proper mass); an employer of the concept of mass is committed to the two laws given above, which are incompatible.[77]

If inconsistent concepts are confused, then for every inconsistent concept, there are two or more *component concepts* that are not being distinguished properly. In the mass example, relativistic mass and proper mass are the component concepts of the inconsistent concept of mass. If truth is an inconsistent concept, and all inconsistent

[76] Perhaps Yablo (1993b) and Eklund (2002) should count, but there are good reasons to be dissatisfied with them (expressed in the previous section); see also Gupta (1999) on misconceptions; see also Field (1973, 1974), Priest (1995), Allen (2001), Camp (2002), and Scharp (2005b) on confused concepts; see also Dummett (1973: 454–5) and Williamson (2003, forthcoming) on defective concepts.

[77] This example is prominent in Field (1973, 1974). See Camp (2002) and Scharp (2005b) for discussion.

concepts are confused, then an important question is: what are the component concepts of truth? I give my answer to this question below; for now it is enough to know that I propose two component concepts for truth.

A second central claim of the theory of inconsistent concepts I offer is that inconsistent concepts should not be employed. There is an important distinction between concept employment and concept possession. I possess many concepts that I do not employ (e.g. the concept of mass). If one employs an inconsistent concept, then one is committed to its constitutive principles, which are incompatible. Because one should avoid undertaking incompatible commitments, one should not employ an inconsistent concept. Instead of employing an inconstant concept, one should employ its component concepts. I call this the *replacement policy* for handling inconsistent concepts. Because I advocate the replacement policy, I often use the term 'replacement concept' for each of the component concepts of an inconsistent concept.

So the two central claims of the theory of inconsistent concepts are: (i) inconsistent concepts are confused; and (ii) one should replace an inconsistent concept with its component concepts in the repertoire of concepts one employs. We can get a general picture of how this account applies to truth. First, truth is a confused concept, which implies that there are components of truth. Second, if there are two component concepts for truth, then we need three theories of truth: one for each component concept and one for our inconsistent concept of truth. Third, we should no longer employ our inconsistent concept of truth. Instead, we should employ the component concepts. Fourth, it follows that the three theories of truth should *not* appeal to our inconsistent concept of truth; indeed, no theory should appeal to our inconsistent concept of truth. Instead, the theory of our inconsistent concept of truth should appeal to the component concepts of truth and the theory for each component should be explained by appeal to whatever other concepts are appropriate (as long as they aren't, in turn, explained in terms of truth).

The theory of our inconsistent concept of truth should explain how we should interpret discourse that expresses this inconsistent concept; in particular it should specify: (i) how we should understand arguments that express truth (i.e. a logic for truth); (ii) how we should understand sentences that express truth (i.e. a semantic theory for truth); and (iii) how we should understand speech acts that express truth (i.e. a pragmatic theory for truth). It is my view that an adequate account of confusion should give us a good idea of how to construct these theories. In particular, I advocate a logic for truth that is a type of relevance logic, a semantic theory for truth that is based on an inferential role theory of meaning, and pragmatic theory for truth that is based on a scorekeeping theory of linguistic usage.[78]

[78] I do not have the space to discuss these theories here; see Scharp (2005a) for details and references.

When considering an inconsistent concept like truth, it is essential to distinguish between several sets of rules for using it. First, there are the incompatible rules that are constitutive of truth (i.e. the truth rules). Those who employ truth *try* to follow these rules. Second, there are the rules stipulated by the logic, the semantic theory, and the pragmatic theory for truth. An interpreter who knows that truth is inconsistent treats those who employ it as if they are bound by these rules. Third, there are the rules stipulating that truth should not be employed at all. Those who know that truth is inconsistent are bound by these rules.

Although there are many complicated issues involved in deciding on the best account of the components of truth, it seems to me that there are there are two fundamental components: *ascending truth* and *descending truth*.[79] Both of them obey versions of mono-aletheism (i.e. no truth bearer is both ascending true and ascending false, and no truth bearer is both descending true and descending false) and substitution (i.e. any singular term that is coreferential with ⟨⟨p⟩⟩ can be substituted in ⟨⟨p⟩ is ascending true⟩ or ⟨⟨p⟩ is descending true⟩ without changing the ascending truth value or the descending truth value). Ascending truth obeys a version of the ascending truth rule for every truth apt truth bearer (i.e. ⟨⟨p⟩ is ascending true⟩ follows from ⟨p⟩), while descending truth obeys a version of the descending truth rule for every truth-apt truth bearer (i.e. ⟨p⟩ follows from ⟨⟨p⟩ is descending true⟩). However, ascending truth obeys a version of the descending truth rule (i.e. ⟨p⟩ follows from ⟨⟨p⟩ is ascending true⟩) for some truth-apt truth bearers, but not all of them. Likewise, descending truth obeys a version of the ascending truth rule (i.e. ⟨⟨p⟩ is descending true⟩ follows from ⟨p⟩) for some truth-apt truth bearers, but not all of them. There are precise ways of characterizing these restrictions, but roughly, if one substitutes a truth predicate for the ascending truth predicates and the descending truth predicates in a sentence and the result is paradoxical for a theory of truth that validates the truth rules, then the original sentence does not obey either the descending rule for ascending truth or the ascending rule for descending truth. Call these sentences *pathological*. It turns out that pathological sentences are ascending true and descending false.

Ascending truth and descending truth are both partially defined in the sense that their extensions and anti-extensions are not exhaustive of the set of declarative sentences; there are ascending truth gaps and descending truth gaps, which are the same. Although I want to permit a range of views on the gaps, there is group of sentences that definitely count as ascending and descending truth gaps: the sentences that express our inconsistent concept of truth. Indeed, it is my view that any sentence

[79] Depending on one's views on the distinction between weak truth and strong truth, one's views on deflationism, and one's views on truth bearers, truth aptness, and truth-values, one might want to distinguish between more than two components of truth. For simplicity, I discuss just two.

that expresses an inconsistent concept is an ascending gap and a descending gap (because they are the same, I will just use 'gap' from here on). The reason is simple. If a sentence has a particular truth-value, then it is acceptable to assert that it has this truth-value. If one asserts that a sentence that expresses an inconsistent concept is true (false), then one is committed to the proposition expressed by the sentence in question (its negation) because of the truth rules. If one is committed to a proposition containing an inconsistent concept, then one is committed to the constitutive principles for that inconsistent concept.[80] Therefore, if sentences that express an inconsistent concept have truth-values, then there is no way to avoid employing that inconsistent concept. Because it is essential to distinguish between concept employment and concept possession when it comes to inconsistent concepts, every sentence that expresses an inconsistent concept is gappy.

I want to turn to the approach to the liar paradox and the revenge paradoxes. Ascending truth and descending truth differ on the pathological sentences. Consider the following two sentences:

(α) (α)is ascending false.
(δ) (δ)is descending false.

These sentences are pathological. However, they are not paradoxical—it is *not* the case that either the theory of ascending truth or the theory of descending truth has the following consequences: (i) (α) is both ascending true and ascending false; (ii) (α) is both descending true and descending false; (iii) (δ) is both ascending true and ascending false, and (iv) (δ) is both descending true and descending false. Indeed, these theories imply that (α) and (δ) are both ascending true and descending false. Consider the analogs of the liar reasoning for (α) and (δ) (with 'AT' in for 'ascending truth' and 'DT' in for 'descending truth'):

(α) is AT (assumption)	(δ) is DT (assumption)
'(α) is AF' is AT (substitution)	'(δ) is DF' is DT (substitution)
* (α) is AF (descending)	(δ) is DF (descending)
(α) is AF (assumption)	(δ) is DF (assumption)
'(α) is AF' is AT (ascending)	* '(δ) is DF' is DT (ascending)
(α) is AT (substitution)	(δ) is DT (substitution)
∴ (α) is AT iff (α) AF	∴ (δ) is DT iff (δ) is DF

Neither of these arguments is valid. In the argument concerning (α), the third step (marked with a '*') is invalid, and in the argument concerning (δ), the fifth step (marked with a '*') is invalid. Of course, one can prove that both (α) and (δ) are AT and DF, but that is not a contradiction.

[80] This is not quite right; I can endorse the proposition expressed by 'acorns are Xnot true' without committing myself to employing the concept of truth, but this sort of example is rare.

The details of how the inconsistency theory works will have to wait for some other occasion, but, roughly, one uses the theory of ascending truth and the theory of descending truth to assign semantic values to the sentences containing 'true' and then one uses a special logic to evaluate the arguments containing these sentences. The semantic values have an epistemic interpretation—they are similar to the 'told true', 'told false', 'told neither', and 'told both' values that are familiar from 4-valued semantics for some relevance logics (except that there are more then just four options in the case of truth). The logic used to evaluate the arguments is a type of relevance logic that is appropriate for confused concepts.[81] Once one has an account of inferential correctness from the logic, one can use an inferential role theory of meaning to assign meanings to the sentences that contain 'true' based on their inferential roles; one can also use a scorekeeping theory of speech acts to assign pragmatic features to utterances of sentences that contain 'true'.[82]

Does the inconsistency theory of truth face revenge paradoxes? I cannot guarantee that it does not. However, given that my analyses of self-refutation problems and inconsistency problems are correct, I can argue that it does not. First, notice that the inconsistency theory of truth does not imply that the constitutive principles for truth are true or valid. In addition, this theory of truth does not employ the inconsistent concept of truth. The logic does not use truth-values and it does not explain validity in terms of truth preservation; instead, it uses epistemically interpreted semantic values and it explains validity in terms of profitability preservation. The pragmatic theory does not explain assertibility in terms of truth; instead, it uses the same resources as the logic. The semantic theory does not assign truth conditions to the sentences in its scope; instead, it assigns inferential roles. Thus, one can accept this theory of truth without employing the inconsistent concept of truth it describes.

My explanation of the inconsistency problems is that they occur for theories of truth that both validate the truth rules and classify some of the paradoxical sentences as defective, where the class of defective sentences does not include all the paradoxical ones. However, the inconsistency theory of truth I endorse classifies all the sentences that express the inconsistent concept of truth as defective. Thus, it does not face any inconsistency problems. The theories of the component concepts of truth do not respect the truth rules; hence they do not give rise to inconsistency problems.

Self-refutation problems occur for theories of truth that validate the truth rules and classify all the paradoxical sentences as defective, including some of their

[81] See Belnap (1976, 1977) for more on the 4-valued semantics and the epistemic interpretation of the semantic values; see also Camp (2002) for the use of this logic on confused expressions. Note that the logic is based on profitability preservation instead of truth preservation. This is essential if one wants to avoid appealing to the inconsistent concept of truth in one's theory of the inconsistent concept of truth.

[82] See Lewis (1979) for an example of this kind of theory.

own consequences. However, the inconsistency theory of truth has no paradoxical sentences as consequences because it does not employ the inconsistent concept of truth. It does not classify the sentences in its scope as true or false. Instead, it implies that all sentences are in the range of inapplicability for the inconsistent concept of truth (i.e. truth has empty extension and empty anti-extension). The theory of the component concepts classifies all the sentences that express the inconsistent concept of truth as gaps. Still, one might wonder whether the theory of ascending truth or the theory of descending truth faces a self-refutation problem. Call the theory of ascending truth T_A and the theory of descending truth T_D. It might seem that both theories are pathological (i.e. they are both AT and DF). Even if this were true, it would not constitute a self-refutation problem because pathological sentences can be acceptable. However, neither theory is pathological. Consider T_A. It implies that (α) is AT and DF. One might be tempted to infer from the claim that T_A implies that (α) is AT, that T_A implies (α). However, the rule, from $\langle\langle p\rangle$ is AT$\rangle$ infer $\langle p\rangle$, is invalid for pathological sentences. Now consider T_D and the following argument:

(i) $T_D$implies that (δ) is AT and DF.
(ii) (δ) is '(δ) is DF'.
(iii) Hence, (δ) is a consequence of T_D.
∴ (iv) T_D is AT and DF.

The problem with this argument is that pathologicality is not preserved by the consequence relation. One can define validity in terms of AT and DT: an argument is *valid* if it preserves DT and the absence of AF (i.e. if the premises are DT, then the conclusion is DT, and if the conclusion is AF, then one of the premises is AF). '(δ) is DF' is a consequence of T_D, and '(δ) is DF' is (δ); hence, (δ) is a consequence of T_D. However, in order to show that T_D is pathological, one would have to show that '(δ) is DT' is a consequence of T_D. However, '(δ) is DT' is not a consequence of T_D because '(δ) is DF' is a consequence of T_D. Therefore, neither T_A nor T_D is pathological.

I want to be clear about my suggestion: I am not expecting people to stop using 'true' and start using 'descending true' and 'ascending true'. It is crucial that people be able to continue using 'true', but because of the division of linguistic labor and their propensity to defer to the experts in cases where it matters, their use of 'true' will no longer express our inconsistent concept of truth, but rather a general concept (i.e. 'true' will mean something like *either descending true or ascending true*).

13.6 Conclusion

I offer a look back and a look forward. First, there are two types of revenge paradoxes: inconsistency problems and self-refutation problems. Second, any theory of truth

that offers an approach to the liar paradox and validates the truth rules is inconsistent, self-refuting, or restricted (that is the conclusion of the revenge argument). Third, truth is an inconsistent concept—a claim justified by three arguments: (i) an inconsistency theory of truth provides the best explanation for revenge paradoxes; (ii) only an inconsistency theory of truth that implies that the truth rules are constitutive but invalid can avoid revenge paradoxes; and (iii) only the particular inconsistency theory of truth I offer has been able to avoid revenge paradoxes. Fourth, the best inconsistency theory of truth treats truth as a confused concept, identifies ascending and descending truth as its components, and does not appeal to our inconsistent concept of truth at all (indeed, it implies that our concept of truth should not be employed).

If these points are correct, then there is much to be done. First, we need to rethink our views on the nature of truth. Deflationists, correspondence theorists, and the rest treat truth as a consistent concept. If truth is an inconsistent concept, then all familiar views on the nature of truth are unacceptable. Second, we need a good theory of inconsistent concepts, which should include a logic, a semantic theory, and a pragmatic theory; if I am right that inconsistent concepts are confused, then a major part of a theory of inconsistent concepts will be a theory of confusion. Third, we need a theory of the replacement concepts for truth (i.e. ascending truth and descending truth), which should include a logic, a semantic theory, and a pragmatic theory. Fourth, we need to use the theory of the replacement concepts for truth and the theory of inconsistent concepts to arrive at a theory of our inconsistent concept of truth. Fifth, we need to start thinking about the nature of the replacement concepts. It seems to me that a broadly deflationist account will work best, but that is little more than an educated guess at this point. Sixth, we need to start thinking about consequences of this account of truth for other concepts that are typically explained by appeal to truth, including: validity, consistency, completeness, necessity, knowledge, meaning, and assertion.[83] Finally, we need to consider whether analogous approaches work for other paradoxes, including: the paradoxes of predication, the paradoxes of reference, the set-theoretic paradoxes, the paradoxes of vagueness, and the paradoxes of the infinite.[84, 85]

[83] I am especially interested in the effects the theory has on the fundamental theorems of mathematical logic (e.g. Gödel's theorems).

[84] My suggestions for how to pursue most of the issues mentioned in this paragraph can be found in Scharp (2005a).

[85] Thanks to Bob Brandom, Hartry Field, Anil Gupta, John McDowell, Stewart Shapiro, Neil Tennant, Matti Eklund, Doug Patterson, Eric Carter, and A. Duncan Kerr for conversations on these issues and comments on earlier drafts.

References

Allen, Martin (2001). 'A paraconsistent-preservationist treatment of a common confusion Concerning predicate extensions'. In John Woods and Bryson Brown (eds.), *Logical Consequence: Rival Approaches* (Proc. 1999 Conf. of the Society for Exact Philosophy) Middlesex: Hermes Science

Atlas, Jay (1989). *Philosophy without Ambiguity*. Oxford: Oxford University Press

Azzouni, Jody (2003). 'The strengthened liar, the expressive strength of natural languages, and regimentation', *The Philosophical Forum* 34: 329–50

Barwise, Jon, and Etchemendy, John (1987). *The Liar: An Essay on Truth and Circularity.* Oxford: Oxford University Press

Beall, JC (1999). 'Completing Sorensen's menu: a non-modal Yabloesque curry', *Mind* 108: 737–9

——— (2001). 'Is Yablo's paradox non-circular?' *Analysis* 61: 176–87

——— (ed.) (2003). *Liars and Heaps: New Essays on Paradox*. Oxford: Clarendon Press

Belnap, Nuel (1976). 'How a computer should think'. In G. Ryle (ed.), *Contemporary Aspects of Philosophy*. London: Oriel Press

——— (1977). 'A useful four-valued logic'. In J. M. Dunn and G. Epstein (eds.), *Modern Uses of Multiple-Valued Logic*, Dordrecht: D. Reidel

Bromberger, Sylvain (1989). 'Types and tokens in linguistics'. In A. George (ed.), *Reflections on Chomsky*. Oxford: Blackwell

Bueno, Otávio, and Colyvan, Mark (2003a). 'Paradox without satisfaction', *Analysis* 63: 152–6

——— (2003b). 'Yablo's paradox and referring to infinite objects', *Australasian Journal of Philosophy* 81: 402–12

Burge, Tyler (1979). 'Semantical paradox', *The Journal of Philosophy* 76: 169–98

Camp Jr, Joseph L. (2002). *Confusion: A Study in the Theory of Knowledge.* Cambridge: Harvard University Press

Cappelen, Herman (1999). 'Intentions in words', *Noûs* 33: 92–102

Chapuis, André, and Gupta, Anil (eds.) (2000). *Circularity, Definition and Truth.* New Delhi: Indian Council of Philosophical Research

Chihara, Charles (1979). 'The semantic paradoxes: a diagnostic investigation', *Philosophical Review* 88: 590–618

——— (1984). 'The semantic paradoxes: some second thoughts', *Philosophical Studies* 45: 223–9

Davidson, Donald (1974). 'Belief and the basis of meaning'. In Davidson (1984)

——— (1982). 'Communication and convention'. In Davidson (1984)

——— (1984). *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press

Dummett, Michael (1973). *Frege: Philosophy of Language*, Cambridge: Harvard University Press

——— (1999). 'Of what kind of thing is truth a property?' In Blackburn and Simmons (eds.), *Truth*. Oxford: Oxford University Press

Eklund, Matti (2002). 'Inconsistent languages', *Philosophy and Phenomenological Research* 64: 251–75

——— (forthcoming). 'The liar paradox, expressibility, and possible languages'. In JC Beall (ed.), *Revenge Paradoxes*. Oxford: Oxford University Press

Feferman, Solomon (1982). 'Toward useful type-free theories I', *The Journal of Symbolic Logic* 49: 75–111

——— (1991). 'Reflecting on incompleteness', *Journal of Symbolic Logic* 56: 1–49

Field, Hartry (1973). 'Theory change and the indeterminacy of reference'. In Field (2001a)

——— (1974). 'Quine and the correspondence theory'. In Field (2001a)

——— (2002). 'Saving the truth schema from paradox', *Journal of Philosophical Logic* 31: 1–27

——— (2003a). 'A revenge-immune solution to the semantic paradoxes', *Journal of Philosophical Logic* 32: 139–77

——— (2003b). 'The semantic paradoxes and the paradoxes of vagueness'. In Beall (2003)

——— (2003c). 'No fact of the matter', *Australasian Journal of Philosophy* 81: 457–80

——— (2004). 'The consistency of the naïve theory of properties', *The Philosophical Quarterly* 54: 78–104

——— (2005a). 'Is the liar sentence both true and false?' In JC Beall and B. Armour-Garb (eds.), *Deflationism and Paradox*. Oxford: Oxford University Press

——— (2005b). 'Variations on a theme by Yablo'. In JC Beall and B. Armour-Garb (eds.), *Deflationism and Paradox*. Oxford: Oxford University Press

——— (forthcoming). 'Solving the paradoxes, escaping revenge'. In JC Beall (ed.), *Revenge Paradoxes*. Oxford: Oxford University Press

Fodor, Jerry, and Lepore, Ernest (1994). *Holism: A Shopper's Guide*, Cambridge: MIT Press

Gaifman, Haim (1992). 'Pointers to truth', *The Journal of Philosophy* 89: 223–61

——— (2000). 'Pointers to propositions'. In Chapuis and Gupta (2000)

Glanzberg, Michael (2001). 'The liar in context', *Philosophical Studies* 103: 217–51

——— (2004). 'A contextual-hierarchical approach to truth and the liar paradox', *Journal of Philosophical Logic* 33: 27–88

——— (2005). 'Truth, reflection, and hierarchies', *Synthese* 142: 289–315

Gupta, Anil (1999). 'Meaning and misconceptions'. In R. Jackendoff, P. Bloom, and K. Wynn (eds.), *Language, Logic, and Concepts*. Cambridge: MIT Press

Gupta, Anil, and Belnap, Nuel (1993). *The Revision Theory of Truth.* Cambridge: MIT Press

Gupta, Anil, and Martin, Robert L. (1984). 'A fixed point theorem for the weak Kleene valuation scheme', *Journal of Philosophical Logic* 13: 131–5

Hardy, James (1995). 'Is Yablo's paradox liar-like', *Analysis* 55: 197–8

Hazen, Allen (1990). 'A variation on a paradox', *Analysis* 50: 7–8

Herzberger, Hans G. (1982a). 'Naïve semantics and the liar paradox', *The Journal of Philosophy* 79: 479–97

——— (1982b). 'Notes on naïve semantics', *Journal of Philosophical Logic* 11: 61–102

Horn, Lawrence (1989). *A Natural History of Negation*. Stanford: CSLI Publications

Horwich, Paul (1998). *Truth.* 2nd edn. Oxford: Clarendon Press

Hugly, Philip, and Sayward, Charles (1981). 'Expressions and tokens', *Analysis* 41: 181–9

Kaplan, David (1973). 'Bob and Carol and Ted and Alice', *Approaches to Natural Language*, Hintikka et Ketland (2004, 2005)

——— (1990). 'Words', *The Aristotelian Society, Supplementary Volume*, 64: 93–119

Ketland, Jeffrey (2004). 'Bueno and Colyvan on Yablo's paradox', *Analysis* 64: 165–72

Koons, Robert (1992). *Paradoxes of Belief and Strategic Rationality.* Cambridge: Cambridge University Press

——— (2000). 'Circularity and hierarchy'. In Chapuis and Gupta (2000)

Kripke, Saul (1975). 'Outline of a theory of truth', *The Journal of Philosophy* 72: 690–716

Leitgeb, Hannes (2002). 'What is a self-referential sentence? Critical remarks on the alleged (non-)circularity of Yablo's paradox', *Logique et Analyse* 177–8: 3–14

Lewis, David (1979). 'Scorekeeping in a language game'. In *Philosophical Papers*, vol. 1. Oxford: Oxford University Press, 1983

Martin, R. M. (ed.) (1984). *Recent Essays on Truth and the Liar Paradox*. Oxford: Clarendon Press

Maudlin, Tim (2004). *Truth and Paradox: Solving the Riddles*. Oxford: Oxford University Press

McDowell, John (1994). *Mind and World*, Cambridge: Harvard University Press

McGee, Vann (1991). *Truth, Vagueness, and Paradox*. Indianapolis: Hackett

Meyer, Robert K., Routley, Richard, and Dunn, J. Michael (1979). 'Curry's paradox', *Analysis* 39: 124–8

Parsons, Charles (1974). 'The liar paradox', *Journal of Philosophical Logic* 3: 381–412

Parsons, Terence (1984). 'Assertion, denial, and the liar paradox', *Journal of Philosophical Logic* 13: 137–52

Patterson, Doug (2006). 'Tarski, the liar, and inconsistent languages', *The Monist* 89

Peacocke, Christopher (1992). *A Study of Concepts*. Cambridge: MIT Press.

Priest, Graham (1990). 'Boolean negation and all that', *Journal of Philosophical Logic* 19: 201–15

_____ (1995). 'Multiple denotation, ambiguity and the strange case of the missing amoeba', *Logique et Analyse*, 38: 361–73

_____ (1997). 'Yablo's paradox', *Analysis* 57: 236–42

_____ (2006). *Doubt Truth to be a Liar*. Oxford: Oxford University Press

Priest, Graham, Beall, JC, and Armour-Garb, Brad (eds.) (2004). *The Law of Non-Contradiction: New Philosophical Essays*, Oxford: Oxford University Press

Prior, A. N. (1960). 'The runabout inference ticket',

Quine, W. V. (1960). *Word and Object*. Cambridge: MIT Press

Scharp, Kevin (2005a). 'Truth and aletheic paradox', Ph.D. diss., University of Pittsburgh

_____ (2005b). 'Scorekeeping in a defective language game', *Pragmatics and Cognition* 13: 203–26

_____ (FTT). 'Fragmentary theories of truth', in preparation

_____ (TI). 'Truth and internalizability', in preparation

_____ (RB). 'Risky business: truth and paradoxicality', in preparation

_____ (AP). 'The aletheic problem', in preparation

Schiffer, Stephen (2003). *The Things We Mean*. Oxford: Oxford University Press

Simmons, Keith (1993). *Universality and the Liar: An Essay on Truth and the Diagonal Argument.* Cambridge: Cambridge University Press

Simons, Peter (1982). 'Token resistance', *Analysis* 42: 195–203

Skyrms, Brian (1982). 'Intensional aspects of semantical self-reference'. In Martin (1984)

Sorensen, Roy A (1998). 'Yablo's paradox and kindred infinite liars', *Mind* 107: 137–56

Stebbins, Sarah (1992). 'A minimal theory of truth', *Philosophical Studies* 66: 109–37

Szabó, Zoltán Gendler (1999). 'Expressions and their Representations', *Philosophical Quarterly* 49: 145–63

Tappenden, Jamie (1999). 'Negation, denial, and language change in philosophical logic'. In D. Gabbay and H. Wansing (eds.), *What is Negation?* Dordrecht: Kluwer

Tennant, Neil (1995). 'On paradox without self-reference', *Analysis* 55: 199–207

Truncellito, David (2000). 'Which type is tokened by a token of a word-type?' *Philosophical Studies* 97: 251–66

Wetzel, Linda (1993). 'What are occurrences of expressions?' *Journal of Philosophical Logic* 22: 215–20

Williamson, Timothy (1997). 'Imagination, stipulation and vagueness'. In E. Villanueva (ed.), *Philosophical Issues 8: Truth*. Atascadero, CA: Ridgeview

_____ (2000). 'Semantic paradox and semantic change'. In A. Kanamori (ed.), *Proceedings of the Twentieth World Congress of Philosophy* 6. Bowling Green: Philosophy Documentation Center

_____ (2003). 'Understanding and inference', *The Aristotelian Society Supplement* 77: 249–93

_____ (forthcoming). 'Reference, inference and the semantics of pejoratives'. In David Kaplan, J. Almog, and P. Leonardi, (eds.), a *Festschrift* for Oxford: Oxford University Press

van Benthem, J. F. A. K. (1978). 'Four paradoxes', *Journal of Philosophical Logic* 7: 49–72

Yablo, Stephen (1985). 'Truth and reflection', *Journal of Philosophical Logic* 14: 297–349

_____ (1993a). 'Definitions, consistent and inconsistent', *Philosophical Studies* 72: 147–75

_____ (1993b). 'Hop, skip and jump: the agnostic conception of truth', *Philosophical Perspectives* 7: 371–96

_____ (1993c). 'Paradox without self-reference', *Analysis* 53: 251–2

14

Burali–Forti's Revenge

Stewart Shapiro

Typical attempts to provide a sufficiently comprehensive account of truth seem to be subject to a phenomenon called 'revenge'. Although there is no agreed upon account of what the phenomenon of revenge is, the idea seems to be that either it is possible to formulate a so-called strengthened version of the Liar paradox that applies to the account in question, or else the framework forbids the expression of something that, intuitively, is not only manifestly expressible, but is needed to formulate or at least motivate the account in the first place. A particularly ironic form of revenge occurs when an account of truth has it that the theory itself is something other than true.[1] Begin with a standard Liar sentence p which is equivalent to 'p is false'. Consider a simple, perhaps simple-minded, account that simply says that sentences like p are neither true nor false. Now consider a sentence q that is equivalent to 'q is either false or neither true nor false' (or just 'q is not true'). This, too, leads to paradox under plausible assumptions. An advocate of the simple-minded theory may avoid this strengthened Liar by insisting that the locution 'neither true nor false' is not expressible in the language. She might hold that, despite appearances to the contrary, the notion of 'neither true nor false' is not expressible at all—perhaps because it is ultimately nonsense. But then the very account of truth under study becomes itself nonsense, by its own lights, since it says that the original Liar is neither true nor false. If our theorist holds, instead, that the notion of 'neither true nor false' is not expressible in the object language L in question, then her account is not sufficiently comprehensive. It is not that we cannot formulate a consistent account of truth for the restricted language, but we need the banned locution in order to motivate the

[1] Thanks to Kevin Scharp and Graham Priest here.

theory. A natural move is to introduce a Tarski-style hierarchy, but this has its own revenge issues.

The other chapters in this volume concern the notion of truth, and whether it is possible to have a comprehensive, revenge-free theory of truth, with a consistent or paraconsistent logic, accepting or rejecting contradictions, etc. This chapter concerns the Burali–Forti paradox, which, I believe, has its own annoying revenge issues. I canvass various ways of dealing with the paradox. It seems that every one of the available options has difficulties, which should lead to its decisive rejection, were it not that the other options fare no better.

14.1 Preliminaries: Burali–Forti Summarized

Define the *field* of a binary relation R to be the property (or class) of being an object that R relates:

$$\lambda x(\exists y(Rxy \vee Ryx).$$

A *well-ordering* is an irreflexive, transitive, binary relation in which every non-empty sub-property (or class) of the field of R has a least element. It is straightforward to define this notion in a second-order language, with no non-logical terminology:

$$\begin{aligned}\mathbf{WO}(R): \forall x\neg Rxx\&\forall x\forall y\forall z((Rxy\&Ryz) \rightarrow Rxz)\&\forall X((\exists x Xx\&\forall x(Xx \\ \rightarrow \exists y(Rxy \vee Ryx)) \rightarrow \exists y(Xy\&\forall z(Xz \rightarrow (z = y \vee Ryz)))).\end{aligned}$$

There is no first-order formula that expresses the property of being a well-ordering, nor is there a first-order formula that defines a particular well-ordering, except in the trivial case in which the field of R has a fixed, finite bound on its cardinality (see Shapiro (1991, chapter 5, §5.1.3)).

Let us define an *ordinal* to be the order-type of a well-ordering. I do not wish to identify ordinals, so conceived, with sets like Von Neumann ordinals, but I also do not wish to reject such an identification out of hand either. We will leave that matter open, as far as possible. The crucial aspect of ordinals is that they are *objects*, and they are denoted by nominalizations like 'the order-type of R'.

The main axiom governing ordinals is a partial abstraction principle:

$$\begin{aligned}\text{(ORD-ABS)}\quad \forall R_1\forall R_2[(\mathbf{WO}(R_1)\&\mathbf{WO}(R_2)) \\ \rightarrow (\mathbf{ORD}(R_1) = \mathbf{ORD}(R_2) \equiv R_1 \simeq R_2)],\end{aligned}$$

where $R_1 \simeq R_2$ is the statement that R_1 is isomorphic to R_2.

Sad to say, (ORD-ABS) is inconsistent. Indeed, let ON be the property of being an ordinal. That is ON(x) if and only if

$$\exists R(\mathbf{WO}(R)\&x = \mathbf{ORD}(R)).$$

Let α and β be two ordinals. As usual, say that $\alpha < \beta$ if there are relations R and S such that α is the order type of R (i.e. $\alpha = \mathbf{ORD}(R)$), β is the order-type of S, and there is an isomorphism between R and a proper initial segment of S.

Suppose that α is an ordinal and that $\alpha = \mathbf{ORD}(\mathrm{R})$. Let S be the restriction of the $<$-relation to ordinals smaller than α:

$$S\beta\gamma \text{ if and only if } \beta \text{ and } \gamma \text{ are ordinals and } \beta < \gamma < \alpha.$$

Cantor showed that S is itself isomorphic to R. By (ORD-ABS), $\mathbf{ORD}(S) = \alpha$. In other words, each ordinal is the order-type of its predecessors, under the given $<$-relation.

The field of the $<$-relation is ON, and it is easy to show that the relation is itself a well-ordering. So let Ω be the ordinal of $<$. We have $\mathrm{ON}(\Omega)$. Now define a relation B as follows: Bxy if and only if $\mathrm{ON}(x)\&\mathrm{ON}(y)\&x < y < \Omega$. Then B is a well-ordering, and let $b = \mathbf{ORD}(B)$. By the foregoing, $b = \Omega$. However, since B is isomorphic to (indeed, is) a proper, initial segment of ON, we have $b < \Omega$. So $\Omega < \Omega$, which contradicts the fact that $<$ is a well-ordering. This is the Burali–Forti paradox.

So if (ORD-ABS) is understood in full generality, it is false. Some well-orderings do not have ordinals. Since an ordinal, as understood here, just is the order-type of a well-ordering, we are forced to the conclusion that some well-orderings do not have order-types.

14.2 The Iterative Conception: ZFC

Allow me to briefly go over the familiar ground of how the situation is handled—or swept under the rug—in Zermelo–Fraenkel set theory. The iterative hierarchy is defined by transfinite recursion, over the ordinals (i.e. order-types of well-orderings). We restrict attention here to the pure iterative hierarchy, and so do not bother with urelements. The ground level, V_0 is the empty set. If α is an ordinal, then $\mathrm{V}_{\alpha+1}$ is the powerset of V_α, and if λ is a limit ordinal, then V_λ is the union of all V_α for $\alpha < \beta$. An object x is a member of the iterative hierarchy V if there is an ordinal α such that $x \in \mathrm{V}_\alpha$.

The custom, of course, is to identify ordinals with von Neuman ordinals, which are transitive sets well-ordered by the membership relation. This is done for convenience. Why have sets and ordinals to talk about when we can just identify the ordinals with certain sets? The identification is especially convenient, since each von Neumann ordinal has the corresponding order-type. So the $<$-relation on ordinals just becomes the membership relation on von Neumann ordinals.

But does the identification make sense? First, how do we know that for every ordinal, in the present sense, there is a corresponding von Neumann ordinal? Well,

we can show, by transfinite induction, that if α is an ordinal, then the corresponding von Neumann ordinal is a subset of V_α and is thus a member of $V_{\alpha+1}$.

What of the converse? We just saw that there are well-orderings that do not have ordinals. Is it at least the case that every von Neumann ordinal has an order-type, an ordinal in the present sense? It seems to me that the thesis that there is an ordinal corresponding to every von Neumann ordinal is a *presupposition* of set theory, and is needed to show that the axioms of ZFC are true of the iterative hierarchy. Suppose, for example, there is an ordinal $\omega + n$, for each natural number n, but that there is no ordinal corresponding to 2ω. Then the transfinite recursion defining the iterative hierarchy can be continued to each $V_{\omega+n}$, but not to $V_{2\omega}$. In effect, V would just be $V_{2\omega}$ (or what $V_{2\omega}$ would be if 2ω existed). The axiom of replacement would entail the existence of the von Neumann ordinal 2ω, but this set would not have an order-type, nor would it be in the iterative hierarchy.

Consider a more or less standard proof of Zermelo's celebrated well-ordering theorem. Given a non-empty set x, we define a one-to-one function from an initial segment of the ordinals into x, in terms of a choice function on the powerset of x. To go from there to the existence of a well-ordering for x, we have to assume that there is no one-to-one function from the ordinals, as presently understood, into any set. This last is perhaps a hidden lemma, in Lakatos's (1976) sense, of Zermelo's proof. It follows from this assumption that there is an ordinal that is the order type of every well-ordered *set*. A fortiori, there is an ordinal corresponding to every von Neumann ordinal.[2] So the ordinals are indeed isomorphic to the von Neumann ordinals, and the common identification of ordinals with von Neumann ordinals thus makes sense mathematically.

Christopher Menzel (1986) has developed a set theory without the presupposition, or hidden lemma, that every well-ordered set has an order type. On philosophical grounds, he insists that ordinals, construed as the order-types of well-ordered sets, are themselves distinct from any sets, and thus from von Neumann ordinals. He defines an iterative hierarchy in which each ordinal is an urelement. In Menzel's set theory, there is a set of all ordinals, but this set has no ordinal. However, there is a powerset of the set of all ordinals, a powerset of that, and so on. There is natural bijection between the ordinals and the von Neumann ordinals, as above, but there is no set of von Neumann ordinals—the iteration does not go that far. So replacement fails. In particular, for Menzel, the axiom of replacement is restricted to sets that do have an

[2] The presupposition that every well-ordered set has an order-type (and thus an ordinal) allows for much of the power of the iterative hierarchy. As we saw, it underlies the theorem that the axiom of replacement holds in iterative hierarchy. The metaphor is that the hierarchy is as 'high' as it is 'wide'. For example, the powerset of ω is a member of $V_{\omega+2}$, the powerset of the powerset of ω is a member of $V_{\omega+3}$, and so on. By Zermelo's theorem, there is a well-ordering of this powerset, and thus an ordinal corresponding to it. So the iteration continues far into the uncountable.

order-type or, equivalently, to pure sets. It seems to me that Zermelo's theorem must also be similarly restricted: we can only show that every set that is equinumerous to a pure set has a well-ordering (and thus an ordinal).

14.3 Of Mice and Set Theorists: Proper Classes, Super-ordinals, and Super-duper Ordinals

By all accounts, both first-order and second-order ZFC seem to be consistent. There is no extended Burali–Forti paradox that can be formulated in the language. Moreover, the theory can be motivated, via the iterative conception, using only notions like that of ordinal, as presently construed, and the ordinary set-theoretic constructions of powerset and union. So far, so good.

On the intended interpretation, the quantifiers of first-order ZFC range over pure sets. These are the only objects, and the only collections, recognized to exist by the theory. But are those the only pure set-like entities that there are? If not, then perhaps our set theory is not sufficiently general.

Proper classes are collections of (iterative) sets that are not themselves sets, because they have members of unbounded rank. Proper classes are not in the range of the variables of first-order set theory. The issue at hand is whether there are any.

Set theorists introduce linguistic items that at least look like singular terms that stand for proper classes. The above 'V' is the accepted term for the iterative hierarchy, the symbol 'Ω' is used for the von Neumann ordinals, and 'L' is the hierarchy of constructible sets, used in Gödel's proof of the consistency of the generalized continuum hypothesis and the axiom of choice. This literary device is not particularly telling. Uses of these locutions are easily paraphrased away, in terms of predicates. For example, '$\alpha \in \Omega$' is just shorthand for 'α is a von Neumann ordinal' and $V = L$ is just the statement that every set is constructible.

The situation is potentially more serious for those who, like Zermelo himself, advocate a second-order version of Zermelo–Fraenkel set theory. Instead of schemes for separation and replacement, one uses single axioms:

$$\forall X \forall x \exists y \forall z (z \in y \equiv (z \in x \& Xz))$$
$$\forall R[\forall x \forall y \forall z((Rxy \& Rxz \rightarrow y = z)) \rightarrow \forall x \exists y \forall z (z \in y \equiv \exists w (w \in x \& Rwz))]$$

See Shapiro (1991, Chapter 5). If the variables X and R here are to be understood along the lines of first-order variables, then each must have a 'range'. An item in the range of X would be a property or some sort of collection-like entity that need not be a set. An item in the range of R would be a binary relation, or a class of ordered-pairs. So

understood, the range of the second-order variables includes proper classes, or things much like proper classes.

George Boolos (1998b, 35–6) objected strongly to such a reading of second-order set theory and, indeed, to any talk of proper classes:

> Wait a minute! I thought that set theory was supposed to be a theory about all, 'absolutely' all, the collections that there were and that 'set' was synonymous with 'collection' . . . If one admits that there are proper classes at all, oughtn't one to take seriously the possibility of an iteratively generated hierarchy of collection-theoretic universes in which the sets which ZF recognizes [merely] play the role of ground-floor objects? I can't believe that any such view of the nature of '$\in$' can possibly be correct. Are the reasons for which one believes in [proper] classes really strong enough to make one believe in the possibility of such a hierarchy?

Boolos's point is that if we are going to recognize something like proper classes, entities that have members, or something similar to members, but are 'too big' to be in the iterative hierarchy, then why not define an iterative hierarchy over one of these proper classes? Think of V as V_Ω, and then continue with $V_{\Omega+1}$, the powerset of V, and then onto $V_{\Omega+2}$, the powerset of that, and 'eventually' to $V_{2\Omega}$, etc. This is just to 'continue' the original iteration that defines the cumulative hierarchy. And, put this way, it sounds silly. To formulate Boolos's point from another perspective, if there is a collection-like thing V that results from the iterative definition, then we have not carried the iteration 'far enough'. If we are going to take the iterative hierarchy seriously, we should think of it as going through *every* well-ordering, and so there is no set-like thing that we 'end up' with, so to speak. From Boolos's perspective, then, there simply are no proper classes. Similar strictures apply to what may be called 'super-ordinals'—which would be the types of well-orderings that are too big to be ordinals—and super-cardinals, which would be the sizes of proper classes. Cantor's 'inconsistent multiplicities'—considered as genuine objects—seem to be the same sort of thing as proper classes, and they are just as illegitimate. Or so the argument goes.[3]

Boolos's argument here is the main motivation behind his well-known pluralist reading of (monadic) second-order quantifiers (see, for example, Boolos (1984, 1985)). Consider, for example, an instance of second-order comprehension:

$$\exists X \forall x (Xx \equiv x \notin x).$$

According to Boolos, this should not be read, 'there is a class (or property) X such that for all sets x, x is a member of (or falls under) X just in case x is not a member of itself'.

[3] In Shapiro (1991) I defined 'logical sets' to be collections taken from a fixed universe. What counts as a logical set is thus sensitive to context. If the universe in question is the iterative hierarchy, then some 'logical sets' are proper classes. See Shapiro (1999) for a (partial) retraction; also Shapiro and Wright (2006, §7). Menzel (1986) is quite happy with large collections, such as the totality of all ordinals, being sets.

Instead, the sentence should be something like 'there are some sets X such that x is one of them just in case x is not a member of itself'. This is an attempt to make sense of second-order set theory, and secure some of the theoretical advantages thereof, without admitting special items for the higher-order quantifiers to range over.

As noted in Shapiro (2003) (and Shapiro and Wright (2006)), however, things are far from comfortable on this score. The problem is in the neighborhood of Burali–Forti. For now, let use the symbol 'Ω' to stand for the *property* of being an ordinal, or, to use the language of plurals, we use 'Ω' to stand for the ordinals, with locutions like '$\alpha \in \Omega$', '$\Omega\alpha$', 'α is an ordinal', and 'α is one of the ordinals' all synonymous with each other. It is harmless here to identify ordinals with von Neumann ordinals—that difference won't matter. So the following can be carried out in ordinary, second-order set theory. We don't need special variables ranging over ordinals.

As noted, the relation of 'less than' on ordinals (or membership on the von Neumann ordinals) is itself a well-ordering: to use the terminology of plurals, given any ordinals, there is an ordinal α such that α is one of them, and if β is also one of the given ordinals, then either $\alpha = \beta$ or $\alpha < \beta$. But, of course, there is no order-type of the ordinals, on pain of contradiction. It is easy to define a two-place predicate that apparently characterizes a well-ordering that is strictly *longer* than Ω. Let α and β be ordinals. Say that $\alpha \prec_1 \beta$ if $\alpha \neq 0$ and either $\alpha < \beta$ or $\beta = 0$. That is, we make the order longer just by putting 0 at the 'end'. This is a routine trick. We can also define a relation that intuitively characterizes a well-ordering *twice* as long as Ω: $\alpha \prec_2 \beta$ if either α is a limit ordinal and β is a successor ordinal, or α and β are both limits and $\alpha < \beta$, or α and β are both successors and $\alpha < \beta$. In $\prec_2$, the limit ordinals come before the successors, and the limit ordinals and the successor ordinals are each isomorphic to the ordinals, according to ZFC anyway. Indeed, here is a well-order that is Ω times as long as Ω: let $< x, y >$ be the ordered pair of x and y. If $\alpha, \beta, \gamma, \delta$ are ordinals, then let $< \alpha, \beta > \prec_3 < \gamma, \delta >$ if either $\alpha < \gamma$ or both $\alpha = \gamma$ and $\beta < \delta$. This is not as far as we can go. There is a predicate corresponding to the 'order-type' of any polynomial involving Ω, essentially reproducing what Cantor proposed with ω.

These long well-orderings, so to speak, can all be formulated as formulas in ordinary, *first-order* set theory, although, as above, we do invoke the resources of second-order set theory to state and prove that they characterize well-orderings. We'll return to this. In third-order set theory, we can define the notion of a *super-well-ordering* to be the property had by some pairs of ordinals if they constitute a well-ordering. A *proper-super-well-ordering* is a super-well-ordering that is at least as long as Ω. In other words, X is a proper-super-well-ordering if there is an order preserving function from the ordinals into a proper initial segment of the Xs. Now define a *super-ordinal* to be the 'order-type' of a super-well-ordering. Of course, super-ordinals are not objects (or at least some aren't), and certainly not sets, but each can be coded as a *plurality* of sets (or a property of sets)—a plurality corresponding to a second-order variable. The

super-ordinals are themselves dutifully well-ordered—as expressed in a third-order language—and one can characterize an intuitive well-ordering that is longer than that of the super-ordinals. If we go to fourth-order set theory, we can characterize the notion of a super-duper-well-ordering, and a super-duper ordinal. Moreover, this iteration of language-orders can be carried into the transfinite, characterizing, for each ordinal α, the language of αth-order set theory, and a corresponding level of α-super-ordinal. Indeed, we can carry the 'orders' of the languages and theories up the level of super-ordinals, and beyond those.

Even if this is all consistent, there must be some sympathy with Boolos's claim that the very *first* iteration, the one we used to characterize the iterative hierarchy in the first place, goes as far as possible. There simply is no 'beyond' that. But, as we saw, one can go 'beyond' without invoking anything like proper classes—set-like or ordinal-like *objects*.

Let us get more down-to-earth—if that is what the ordinary iterative hierarchy, treated with a first-order or at most second-order language, is. Transfinite recursions and inductions turn on well-orderings. It might seem that it is legitimate to do such recursions and inductions over ordinals and, it might seem, only over ordinals. On occasion, however, set theorists at least seem to invoke transfinite recursions and inductions whose 'length' is at least that of $\prec_2$, i.e. twice as long the ordinals. For example, the concept L of being a constructible set is defined by transfinite recursion over all ordinals (i.e. of length Ω). But set theorists go on to do transfinite recursions on L, which are also of length Ω. So, in effect, we have a transfinite recursion of length 2Ω.

The constructions invoked in mouse theory make for an interesting case study here. The following appears in a survey article, by Ernest Schimmerling:[4]

> We begin by constructing L level by level. The first ω levels are exactly the hereditarily finite sets, the next ω_1^L levels are exactly the sets that are hereditarily countable in L, and so on. Now we ask ourselves what comes next.
>
> (Schimmerling (2001, 486–7))

Of course, this talk of 'construction' is only a metaphor. What is literally true is that we *define* the constructible sets (L) by transfinite recursion over all ordinals. And proofs about L invoke transfinite induction over all ordinals. But what is the literal meaning of Schimmerling's question, 'what comes next'? What can possibly come *after* the ordinals? He continues:

> For although we have climbed up to the minimal transitive proper class model of ZFC, foundational considerations that fall under the category of *large cardinals* have tempted us to adopt certain theories that extend ZFC. These extensions are not true in L, for they imply

[4] Thanks to Tim Bays for drawing my attention to this arcane branch of set theory.

that there exists a non-trivial elementary embedding j:L $\rightarrow$ L, which is known to fail in L. So how do we continue or revise the construction in a way that buys us the existence of such an embedding? One naïve idea is to continue the construction past all the ordinals and throw in the proper class j at stage Ω or beyond, but this approach leads to some obvious metamathematical problems that we find annoying.

There are also some annoying conceptual problems here, too. The metaphorical text sounds like a lapse into nonsense. If there is a 'past all the ordinals' to 'go on to', we have not gone through all of the ordinals—through *all* of the well-ordering types. Is there a coherent literal reading we can bring to this?

Typically, the way around the 'annoying' meta-mathematical problems to which Schimmerling refers is to replace the long transfinite recursions with codings. That is, the set theorist works hard to simulate what would be the result of a long transfinite recursion within ordinary, first-order set theory. This takes care of the conceptual problem as well. With the coding, there is no lapse into nonsense.

Nevertheless, it seems to me that the grand transfinite recursion is coherent as it stands, or at least as coherent as anything else in set theory. If we can indeed legitimately and intelligibly talk about all ordinals, then, as we saw above, it is straightforward to define a predicate that characterizes a relation of order-type 2Ω. The pairs of ordinals that satisfy this predicate also satisfy the second-order predicate of being a well-ordering. So why can't we do transfinite recursions and inductions over that relation, or over those pairs of ordinals? We can even do a transfinite recursion along the order-type $\prec_3$ above, of length Ω^2. And of course, that is not as far as we can go.

Of course, one must be careful how things are put, to avoid the obvious contradiction. The recursion and induction is done over a proper class, or a relation, or a predicate, or the pairs of ordinals that satisfy the predicate—depending on how one likes to read second-order quantifiers. The 'pain' of contradiction comes if we think of this predicate as defining an order-type, a 'length', or any other sort of ordinal-like *object*. So the talk of the 'length' of the recursions and inductions—2Ω, Ω^2, etc.—is only a metaphor, a manner of speaking. Let us just say that the legitimacy of the technique of long transfinite recursions and inductions is a working hypothesis. What reason is there to demur from it?

At this point, I can almost hear Boolos protesting, paraphrasing his complaint against proper classes in (1998b, 35):

> Wait a minute! I thought that set theory was supposed to include a theory about all, 'absolutely' all, the well-orderings and transfinite recursions that there are, and that 'well-ordering-type' was synonymous with (or at least coextensive with or isomorphic to) 'ordinal'.

The problem here is that predicates corresponding to these 'order-types' are definable—in the *first-order* language—as soon as we make the assumption that we can

talk about, and thus have bound variables ranging over, all ordinals. And what is to prevent that? It is manifest that the predicates do characterize well-orderings. So why can't we do transfinite recursions and inductions on them?

Shapiro (2003) reluctantly delimits what is, admittedly, a thin straw to grasp, in order to reign in Burali–Forti and keep things from blowing up. The key observation is that the definition of a property (or predicate or plurality) being a well-order is second-order (see Shapiro (1991, §5.1.3)). So one can avoid the whole problem by refusing to use second-order variables in theories whose first-order variables do not range over a set. Then no notion corresponding to a 'class-sized well-ordering' can be formulated. But there will also be no formula that expresses the seemingly patent fact that the ordinals are well-ordered. This proposal would block the long transfinite recursions and inductions mentioned above, at least if the text is taken literally.

It is a damn thin straw, though, and pretty uncomfortable to hang on to, especially for me. Even if we put aside the compelling arguments in favor of second-order languages in Shapiro (1991), Boolos's later writings (e.g. (1984), (1985), (1985a)), and elsewhere, the general point remains that denying the existence of the long well-orderings Ω, 2Ω, Ω^2 merely seems like an ad hoc maneuver. To grasp the straw is to claim that the formal language of set theory cannot express what can very well be expressed informally. This, as noted at the outset, is a type of revenge. The constructions of predicates defining the long well-orderings are all first-order. And as noted, one can define the long well-orderings (if that is what they are) as soon as the notion of 'ordinal' has been defined. Moreover, there is no problem of *proving*, intuitively, that the formulas in question characterize well-orderings. Since the relations in question are clearly transitive, the only way they can fail to be well-orderings is if they are not well-founded. Consider, for example, the von Neumann ordinals. Suppose that there were a descending ω-sequence of von Neumann ordinals $\alpha_1, \alpha_2, \alpha_3, \ldots$, where $\alpha_2 \in \alpha_1$, $\alpha_3 \in \alpha_2$, etc. By transitivity, $\alpha_2, \alpha_3, \ldots$ are all in α_1, which contradicts the fact that α_1 is a well-order, under membership. This last is a perfectly good first-order statement. Indeed, the only thing the straw prevents us from doing is to *state* that Ω, for example, is as well-order, even though we can clearly see that it is. As with the revenge of the Liar, here we see a 'solution' that claims that a certainly patently expressible notion is not, in fact, expressible. And this very notion—that of well-ordering—is needed to get the whole project going in the first place.

14.4 Zermelo: Codifying the Upward March

The set theorist can—and perhaps must—claim that no *objects* correspond to the explicitly definable long well-ordering predicates. These predicates simply have

no associated order-types. The grounds for Boolos's rejection of proper classes must also preclude order-types Ω, 2Ω, Ω^2, and the like, provided that these are construed as objects. Perhaps we can still say that such well-orderings exist in some sense, maybe as pluralities, and perhaps we can do the long transfinite inductions and recursions. It is just that there is no 'it' that corresponds to, say 2Ω.

But what, exactly, is wrong with the long transfinite recursions and inductions if we take them literally, as invoking an order-type? Why can't we just introduce such 'ordinals', or names for them, by suitably *expanding* our ontology? We just introduce a singular term, like 'Ω', that is to denote the order-type of all ordinals, without intending to paraphrase it away. This gives rise to another singular term for 2Ω, another for Ω^2, and one for Ω^Ω, and off we go, via the axioms of set theory. In doing so, we are just giving names to well-orderings that we are capable of understanding and using, and treating those well-orderings as objects. The resulting theory is consistent if standard set theory together with an axiom asserting the existence of a strongly inaccessible cardinal is. Moreover, the envisioned theory seems to be *true* when Ω is interpreted as the order-type of the ordinals of ordinary set theory. So what's wrong?

To repeat, what is wrong here is that we have contradicted the understanding of the iterative hierarchy, with which the process starts. The original transfinite recursion, that defines V, was supposed to go through the series of *all* ordinals—all possible well-order-types—not just those that come before the first 'proper' ordinal, or super-duper-ordinal, or whatever. The original transfinite definition was to go over *all*, absolutely all, well-orderings. By introducing new objects at the end (so to speak) we seem to end up with a consistent formal theory, but it does not sustain the intended interpretation of ordinary first set theory.

Maybe the problem was in the attempt to run a transfinite recursion over all ordinals. Maybe there just is no such 'all', whether that is to be interpreted as a singular, as plural, as a property, or whatever. Zermelo (1930) delimits a program for set theory that might be used to sanction a thesis like this. He begins with a version of second-order ZFC with urelements, in pretty much its contemporary form. For present purposes, we will not bother with urelements, and stick to so-called pure set theory. Each model of pure second-order ZFC is isomorphic to a rank V_κ, in which κ is a strong inaccessible.

Zermelo (1930, 1233) proposes the existence of 'an unbounded sequence' of (standard) models of the theory, each larger than its predecessors. Call this the *principle of extendibility*. To speak roughly, the principle entails that the strong inaccessibles are unbounded in the universe. To speak less roughly, the principle is that for each strong inaccessible, there is a larger one.

According to Zermelo, each model M of set theory, so construed, has subsets (like the collection of ordinals in M) which are not members of M. *Within* the given model M, these subsets are proper classes. However,

[w]hat appears as an 'ultra-finite non- or super-set' in one model is, in the succeeding model, a perfectly good, valid set with both a cardinal number and an ordinal type . . . To the unbounded series of Cantor ordinals there corresponds a similarly unbounded . . . series of essentially different set-theoretic models.

Zermelo's ideas are not uncommon among contemporary set theorists, or at least those with realist tendencies. Although, so far as I know, Zermelo does not say so explicitly, perhaps we can deal with Burali–Forti by invoking a sort of systematic ambiguity. The idea is that at any given time, the set theorist thinks of herself as working within (or talking about) an arbitrary, but fixed, model of set theory. The 'proper classes' that she refers to in this talk are actually sets in succeeding models. Let us call this the *Zermelo program*.[5]

Zermelo (1930, 1233) writes:

Scientific reactionaries and anti-mathematicians have so eagerly and lovingly appealed to the 'ultra-finite antinomies' in their struggle against set theory. But these are only apparent 'contradictions', and depend solely on confusing *set theory itself* . . . with individual *models* representing it . . . The two polar opposite tendencies of the thinking spirit, the idea of creative *advance* and that of collection and *completion*, ideas which also lie behind the Kantian 'antinomies', find their symbolic reconciliation in the transfinite number series based on the concept of well-ordering. This series reaches no true completion in its unrestricted advance, but possesses only relative stopping-points, just those [strong inaccessibles] which separate the higher model types from the lower. Thus the set-theoretic 'antinomies', when correctly understood, do not lead to a cramping and mutilation of mathematical science, but rather to an . . . unsurveyable unfolding and enriching of that science.

The Zermelo program nicely handles long transfinite recursions and inductions, such as those invoked in the initial presentation of mouse theory. The idea is that we think of the recursions and inductions as restricted to (or as 'taking place within') a fixed arbitrary model M of set theory. The 'totality' of von Neumann ordinals in M is, of course, not a member of M, but the ordinals in M do constitute a set, and thus a von Neumann ordinal, in later models in the hierarchy (thanks to extendibility). Later models contain von Neumann ordinals with order-types much longer than the totality of von Neumann ordinals of M. The 'long' transfinite recursions and inductions 'take place', so to speak, in a later model, but they are about, and valid for, the given model M.

Zermelo proves a theorem that entails that, given any two models of pure set theory, one of them is isomorphic to an initial segment of the other. This shows that statements whose quantifiers are restricted to a certain accessible rank are true in all of the models or in none. Examples include the continuum hypothesis and just about

[5] The Zermelo program seems to be the natural result of applying the account of quantification sketched by Michael Glanzberg (2004) to set theory.

every statement of real analysis, complex analysis, functional analysis, or the like. This gives the Zermelo program a certain stability. Recall that the theme is that we think of the set theorist as talking about an arbitrary, but unspecified model of set theory. Zermelo's theorems indicate that in a wide variety of cases, it does not matter which model it is.

The viability of the Zermelo program turns on the principle of extendibility, which we formulated neutrally as the thesis that for each strong inaccessible, there is a larger. Zermelo later proposes a stronger principle, postulating 'the *existence of an unbounded sequence of* [strongly inaccessible ranks] as a new *axiom* of "meta-set theory" '. In effect, this meta-axiom states that for each ordinal α, there is a unique strong inaccessible κ_α. It follows that the sequence of standard models is itself equinumerous with the ordinals.

What are we to make of Zermelo's own language, the language in which the structure of models is described, the meta-axioms given, and the aforementioned theorems proved? That is, what are we to make of Zermelo's talk of 'models', 'normal domains' (i.e. strongly inaccessible ranks), 'order types', and the like? Clearly, this language is not intended to be about an arbitrary model of set theory. Rather, it is about the hierarchy of models itself. Indeed, the theorems, principles, and meta-axioms are there to describe the program, and show that it is viable.

Consider, for example, Zermelo's theorem that given any two models of the pure second-order theory, one is isomorphic to an initial segment of the other. As it happens, this is true in each model of second-order ZFC, but that much is not relevant to the dialectic. Moreover, the principle of extendibility, and the meta-axiom that the strong inaccessibles are equinumerous with the ordinals, are *not* true in each model of second-order set theory. Extendibility is true only in ranks that are limits in the series of strong inaccessibles, and the meta-axiom is true only in ranks that are fixed points in the series of strong inaccessibles:[6] cardinals $\kappa_\alpha = \kappa$ in which $\alpha = \kappa$.

This an almost classic instance of revenge. The Zermelo program says that we do not talk about *the* iterative hierarchy; rather we take ourselves to be talking about a fixed, but arbitrary model of the theory. But we cannot describe the program in the first place without violating the restriction it is meant to sanction. That is, in order to get the Zermelo program going, the theorist describes *the* series of models of second-order ZFC, and to make the program work—to show how we can have our cake and eat it too—the theorist proves theorems about this hierarchy, the very hierarchy that we are trying to get away from.

There is at least a vague analogy with the Tarskian hierarchy. On some accounts, we are told that there is no single predicate or property of truth. Rather, there is a

[6] Of course, it is useful to consider models of extendibility and of Zermelo's meta-axiom. The latter suggests another meta-axiom: the fixed points in the series of strong inaccessibles are themselves equinumerous with the ordinals. We are off and running on the topic of so-called small large cardinals.

hierarchy of such predicates, and any given legitimate use of truth invokes a level in the hierarchy. But in order to set up the program, the Tarskian describes and uses a language that has, or at least can refer to, all of the predicates in the hierarchy, and we have no truth predicate for that language (unless we start another hierarchy). We can make the analogy between the Zermelo hierarchy and the Tarski hierarchy tighter if we extent the latter into the transfinite.

Getting back to the matter at hand, perhaps we can combine the Zermelo program with the thin straw delimited at the end of the previous section. Recall that on the program, the language of *ordinary* set theory can be (and indeed is) second-order, since that language is taken to be about a given fixed, but arbitrary model of ZFC. So construed, the second-order quantifiers range over sets in 'later' models in the hierarchy. We encounter a revenge problem only when we come to be talking about the hierarchy of models itself. Here is were we (try to) grasp the thin straw: we take the language used to talk about the hierarchy of models to be first-order.

This seems to block Burali–Forti, as it did in the previous section. We need not worry about an order-type for the totality-of-ordinals-in-all-models (or, with the meta-axiom, an order-type of the models themselves), since with first-order resources, we cannot *state* that the ordinals (or the models) are well-ordered. Indeed, we cannot talk about them—all of them—at all, despite the fact that I seem to be doing so in this paragraph, and in this section generally.

On this plan, we cannot do long transfinite recursions and inductions on the entire hierarchy (so to speak), but perhaps we can live with that. It is getting clear, at least to me, that we are not going to get all that we want. Perhaps grasping the thin straw in a Zermelo-style meta-theory is about as close as we are going to get.

I am not sure how plausible this plan is. How well can we do in describing the hierarchy with a first-order language? Early in (1930) Zermelo explicitly writes that we do apply ordinary set-theoretic concepts to the models themselves:

> . . . we call a 'normal domain' a domain consisting of 'sets' and 'urelements' which satisfies the [ZF] system with regard to the 'basic relation' $a \in b$. We will treat 'domains' of this kind, their 'elements', their 'subdomains', their 'sums' and 'intersections' exactly like *sets*, and thus according to the general *set-theoretic concepts and axioms*, for there is no means of distinguishing them from sets in any way which essentially matters. However, we will always denote them as 'domains' and not as 'sets' in order to distinguish them from the 'sets' which are the elements of the domain in question. (second emphasis mine)

The 'set-theoretic axioms' in question here include separation and replacement, both of which are second-order in Zermelo's formulation. So if we are to take this talk literally, we will have to countenance second-order quantification over the hierarchy of models, after all. We are in danger of giving back all the gains made by the program.

We thus need first-order surrogates for separation and replacement, as applied in the meta-theory to the hierarchy of models (so to speak). Separation is not problematic.

Suppose that x is a set or 'domain' and we wish to use separation on x using a property, formula, or plurality P. By extendibility, x is a member of an strongly inaccessible rank M. The next strong inaccessible after M contains every member of the full powerset of the domain of M, and so it contains a *set* y whose members are the P's in M. The set promised by this instance of separation is just the intersection of x and y. In other words, we only need to apply separation within a given model, and there full second-order resources are unproblematic (thanks to extendibility).

Replacement is not as straightforward. The full power of the extendibility principle and the meta-axiom, and some obvious extensions thereof, does seem to invoke replacement in the meta-theory (see Shapiro (2003, §4)). However, much of this power can be had with a first-order axiom *scheme*, along the lines of the replacement scheme in ordinary, first-order set theory. That is, in the meta-theory, we take, as axioms, each formula obtained from the replacement axiom by replacing the second-order variable with a formula from the first-order language. Again, this falls prey to the arguments against schemes in Shapiro (1991, Chapter 5) and elsewhere (but see Feferman (1991)).

Well, once again, it does not seem plausible to have everything we want and still avoid Burali–Forti. Perhaps the tradeoffs here are not so bad. Still, let us keep trying a bit longer.

14.5 Nominalism: Trying to Do Without

I must confess that in the nearest possible world in which my counterpart is a nominalist, it is the Burali–Forti paradox that drives him there. If there are no abstract objects, then there is no need to worry about the order-type of all ordinals. According to the nominalist, there are no order-types, and thus no ordinals, at all. End of paradox. End of revenge.

Or is it? What of set theory? Presumably, we do not want to drive that out too. Charles Parsons (1977) and Geoffrey Hellman (1989, Chapter 2), (2002) provide accounts of set theory similar to Zermelo's, but they think in terms of *possible* collections, rather than thinking in terms of an actual structure or hierarchy of sets.[7] Recall Zermelo's principle of extendibility: for each strongly inaccessible rank, *there is* a larger. In place of this Hellman (1989, 72) has a modal assertion. In formal mode, it asserts that, necessarily, for every model $< X, f >$ of second-order set theory (and so for every

[7] Hellman (2002) interprets Zermelo's (1930) account of set theory in modal terms, by inserting boxes and diamonds into the text at crucial places. This is more in the way of rational reconstruction than exegesis.

strongly inaccessible rank), there can be another model $< Y, g >$such that X is a proper subclass of Y and f is the restriction of g to X. In symbols:

$$\text{(MOD-EXT)} \qquad \Box\forall X\forall f[(\wedge \mathrm{ZF}^2)^X[\in /f] \rightarrow \Diamond\exists Y\exists g((\wedge \mathrm{ZF}^2)^Y[\in /g]\&(X, f) < (Y, g))],$$

where '$\wedge \mathrm{ZF}^2$' is the conjunction of the axioms of second-order set theory; '$(\wedge \mathrm{ZF}^2)^X[\in /f]$' is the restriction of those axioms to the (monadic) second-order variable X, substituting the binary relation variable f for the membership symbol '$\in$'; and $(X, f) < (Y, g)$ says that $\forall x(Xx \rightarrow Yx)\&\exists y(Yy\&\neg Xy)$ and that f is the restriction of g to X. So (MOD-EXT) is a sentence in a modal second-order language, with no non-logical terminology. One can also formulate a modal version of Zermelo's meta-axiom: necessarily, if α is any ordinal, then it is possible for there to be a structure M satisfying second-order ZFC such that there are α-strong inaccessibles in M. Stronger meta-axioms, along the same lines, are also forthcoming.

The Parsons–Hellman program is an attractive resolution of the paradoxes of set theory, at least for a nominalist. One major problem for this approach is to make sense of the modality involved (see Shapiro (1993)), but considering that would take us too far afield.

Is there a revenge issue? Although Hellman does not reify possible well-orderings, of course, the theory gives us the resources to talk about which well-orderings are possible. Instead of the existence of possibilia, the modal set theorist speaks of the possible existence of certain kinds of objects ordered a certain way. As we have just seen, Hellman's theory makes bold assertions about what possible well-orderings there can be, along the same lines as Zermelo's bold assertions about what well-orderings (and what von Neumann ordinals) there are. Now, any possible collection or plurality of well-orderings is itself well-ordered, under the usual relation delimited above. So let us at least try to consider the possible well-ordering Π of all possible well-orderings. The alleged possible well-ordering of all possible well-orderings can be extended, via the very definition of well-ordering (and thus without invoking anything as powerful as (MOD-EXT)). So Π is not the well-ordering of all possible well-orderings after all. Put otherwise, once we realize that any possible well-ordering can be extended, then there simply is no possible well-ordering of all possible well-orderings. A neat resolution of the Burali–Forti paradox.

But surely we can *formulate* the notion of a possible well-ordering, and we do have quantification at our disposal. So what prevents us from formulating the Burali–Forti paradox in the modal language? And why can't we do transfinite recursions and inductions over *all* possible well-orderings, as in the construction of the iterative hierarchy, or the long transfinite inductions recursions and inductions?

Hellman invokes the now common metaphor of possible worlds to help explain and motivate his modal claims, but he does not take such talk at face value. We follow

suit here, for the time being. The connection with the Zermelo program is striking. Hellman allows second-order quantification *within* each world. This is the counterpart of Zermelo's use of second-order separation and replacement within each model (i.e. each strongly inaccessible rank). For Hellman, the 'proper classes' in each world are sets in another, larger possible world. For Zermelo, the proper classes of each model are sets in other, larger models. For Hellman, there is no world that houses every possible ordinal, or every possible set. And for Zermelo, there is no single model of second-order ZFC that contains an isomorphic copy of all such models.

Some important distinctions play themselves out nicely in the formalism, in terms of the metaphor of possible worlds. In a sense, Hellman's 'quantifier' $\Box\forall x$ covers all objects in all worlds: a formula in the form $\Box\forall x\Phi$ says that Φ holds of all objects in all worlds. So in Hellman's language, we can simulate something like unrestricted first-order quantification: quantification over absolutely all (possible) sets. But second-order quantification is different. If the variable X is monadic, the locution $\Box\forall X$ covers the proper (and improper) classes of *each world*, or the proper (and improper) classes of all worlds. In effect, a formula in the form $\Box\forall X\Phi$ says that in each world w, Φ holds of the classes in w. But remember that each such proper class is a set in another world. So, to continue the metaphor of possible worlds, the locution $\Box\forall X$ also only 'ranges' over possible *sets*.

For Hellman, then, there need be no such things as what may be called 'absolutely proper classes': possible collections, or collection-like things (like V or Ω) that are 'too large' to be in any one world. It is important to ban them since we just saw that there cannot be an ordinal of all possible well-orderings. But how do we accomplish this ban? The comprehension scheme is a staple of second-order logic. To focus on the monadic case, in standard deductive systems, each instance of

$$\exists X\forall x(Xx \equiv \Phi),$$

in which the variable X does not occur free in Φ, is an axiom. Let $O(x)$ be a formula asserting that x is a von Neumann ordinal. The following is thus an instance of comprehension:

$$\exists X\forall x(Xx \equiv \Diamond O(x)).$$

To ease exposition, let us introduce a predicate letter Ω for the promised X of this formula. So Ωx if and only if it is possible that x is a von Neumann ordinal. As usual, the Ωs are themselves well-ordered, under the membership relation: any plurality of Ωs have a least element. So why don't the Ωs themselves have an order type, over which we can do transfinite recursions and inductions (and so off we go)? In effect, the Ωs are (or would be) an 'absolutely proper class'. There are too many for them to be in any one world. Again, such totalities, pluralities, or whatever, are banned.

The rejection of (the analogue of) absolute proper classes motivates a restriction on the *syntax* of the formal language: a formula is well-formed only if it does not contain a sub-formula which contains a free second-order variable within the scope of a modal operator.[8] So the above instance of comprehension is ill-formed.

Hellman thus adopts the analogue of the flimsy straw broached at the end of each of the previous two sections. There is no problem with second-order quantification *within* each world. For example, if Φ is a theorem of ordinary second-order set theory (and so has no modal operators), then $\Box\Phi$ holds in Hellman's system: Φ is true in all worlds. But we cannot use unrestricted second-order formulas when we wish to make *modal* claims about sets and possible sets. The semantic value (or values) of a second-order variable, in this context, would be a banned absolutely proper class.

Just as with the Zermelo program, however, there is a deep need for a replacement principle in Hellman's theory—in the language we are to *use* to describe the modal system. Suppose, for example, that we have convinced ourselves that, necessarily, if n is a natural number, then it is possible for there to be n distinct strong inaccessibles (or n supercompact cardinals, or whatever). A modal version of replacement would allow us to conclude, as we should, that it is possible for there to be ω-many distinct strong inaccessibles (or supercompact cardinals). However, as noted in the previous section and as argued extensively elsewhere (Shapiro (1991, chapter 5)), replacement is best seen as second-order. It has a variable ranging over all (possible) functions. The appropriate modalized version would say that necessarily, if R is a functional relation then, for each set x, it possible for there to be a set y which contains all and only the values of this function as applied to x. Formally,

$$\Box\forall R[\forall x\forall y\forall z((Rxy\&Rxz \rightarrow y = z)) \rightarrow \forall x\Diamond\exists y\forall z(z \in y \equiv \exists w(w \in x\&Rwz))].$$

This, however, violates the restriction on the syntax noted above, since the variable R occurs within the scope of the diamond in the consequent. The function in question seems to represent one of the dreaded absolutely proper classes.

Thus, in setting up the modal system, Hellman does not have a second-order replacement axiom. Instead, he proposes a replacement scheme (Hellman (1989, 78)), along the lines of ordinary first-order set theory—just as we suggested, tentatively, for the Zermelo program.

In sum, then, Hellman has a straightforward and direct analogue of unrestricted first-order quantification, via the locution $\Box\forall$x. But there is, and can be, no analogue of unrestricted second-order quantification, and thus no possible property or plurality of all possible well-orderings.

[8] The relevant clause in the formation rules would be something like this: if Φ is a well-formed-formula that does not contain any free second-order variables, then $\Box\Phi$ and $\Diamond\Phi$ are well-formed formulas.

Is there any revenge here? Notice that in Hellman's system, one can do long transfinite recursions and inductions *within* any given possible world. The well-ordering used in the recursions and inductions corresponds to a von Neumann ordinal in another possible world. In other words, we study the properties of a given well-ordering by considering even longer well-orderings: we study what happens in a given world by focusing attention on another world. So far so good. But why can't we do transfinite recursions and inductions over all possible well-orderings? Simply because we cannot *say* what we need to say to get this started.

Again, let $O(x)$ say that x is a von Neumann ordinal. Then, in effect, a formula in the form $\Box\forall x(O(x) \rightarrow \Phi)$ says that all possible von Neumann ordinals have the property expressed by Φ. We can state and prove that the membership relation on the von Neumann ordinals is transitive:

$$\Box\forall x\forall y\forall z((O(x)\&O(y)\&O(z)\&x \in y\&y \in z) \rightarrow x \in z).$$

And we know that the membership relation is well-founded in general, and thus well-founded on the possible von Neumann ordinals. Intuitively, one would think, this entails that the membership relation on the possible von Neumann ordinals is itself a (possible) well-ordering. The only reason we cannot get Burali–Forti going in the system is that the aforementioned restrictions on comprehension prevent us from *saying* in the formal object language, that the possible well-orderings are themselves well-ordered. But why can't we say it, if we can see it? Moreover, it seems that we just did say that the possible well-orderings are well-ordered. It is just that we cannot say this in the formal language.

The situation with Hellman's modal set theory thus looks analogous to that of the Zermelo program of the previous section. There will be a difference, perhaps, if the indicated restrictions on the modal set theory can be independently motivated in a way that does not provide a motivation for a similar restriction on the Zermelo program. Consideration of that would take us too far afield (see Shapiro (1993)).

Even if this strategy fails, the flimsy straw introduced at the end of the two previous sections may not be quite as flimsy here. Both the Zermelo program and the modal program seem to demand that certain intuitively expressible things are in fact inexpressible. We noted at the outset of this study that the inability to say certain things in a proposed reaction to paradox counts as revenge if one needs to say these things in order to set up the program in the first place.

Recall that Zermelo speaks of an indefinitely extendible sequence of models of set theory. Seemingly, in discussing the program, we talk about *all* such models, and this suggests that we can talk about *all* von Neumann ordinals. Second-order resources would give us something that looks suspiciously like an order-type for them. The analogue in the Hellman program is the talk of the possibility of von Neumann ordinals. This is encouraged by the talk of possible worlds, where we at least seem to

talk about possible objects. If we can talk about 'all possible worlds' in one breath, then why can't we similarly talk about all possible von Neumann ordinals, and their order-type? In the informal gloss, we seem to have stumbled onto absolutely proper classes after all. One response is to simply say that to talk of such things leads to contradiction. To adapt something that Michael Dummett (1991, 316) says in a related context, this response 'is to wield the big stick, but not to offer an explanation'. It is ad hoc.

But perhaps the two programs are not on a par here (*pace* Shapiro (1993)). As a nominalist, Hellman does not reify 'possible von Neumann ordinals', and the talk of possible worlds is only a metaphor, a manner of speaking. The apparent talk of possibilia is just a picturesque way of making modal claims, propositions expressed in the language of set theory augmented with boxes and diamonds. So, officially, we do not need to invoke anything like absolutely proper classes to motivate or set up the program. We just need to make modal claims like (MOD-EXT).

On the other hand, recall that we are kept out of trouble, intuitively, by the restriction on the syntax of the modal language: free second-order variables cannot occur in the scope of a modal operator. This restriction is nicely motivated via the metaphor of possible worlds. Absolutely proper classes—collection like-pluralities that are too many to fit into a single possible world and the possible well-ordering of all possible well-orderings—do not exist simply because any collection-like thing, any ordinal, and any well-ordering can be extended. But the talk of possible worlds is only a metaphor. What is needed, then, is a motivation of the restriction on the syntax that is stated directly in the modal language. Without this, one can be forgiven for charging that the restrictions in question are ad hoc.

I propose to leave the dialogue at this juncture, hopefully with the burdens on each side a bit clearer than they were before.

14.6 Getting Desperate: Embracing (or Almost Embracing) the Contradiction

Graham Priest (1987) has long argued that consistent resolutions of the semantic paradoxes are all subject to revenge. He proposes that we just embrace the contradictions. Priest holds that the argument leading to the contradiction of the Liar, for example, is valid and it has true premises. The Liar sentence is thus both true and not true. If we go this route, of course, we have to adopt a paraconsistent logic, to keep from concluding that every sentence is true (and not true). That is, we have to give up the principle of ex falso quodlibet: from Φ and $\neg\Phi$, infer Ψ.

Priest makes the same claims about the paradoxes of set theory, and Burali–Forti in particular (Priest (1987, chapters 2, 10), (2002, chapters 2, 8), (2006)). As above, we define an ordinal to be the order-type of a well-ordering. We notice that the ordinals are themselves well-ordered, and so there is an ordinal Ω of all ordinals. Since we are taking on the pain of contradiction, we just say this, without fancy paraphrase. We have that $\Omega < \Omega$, since Ω is itself an ordinal. But we also have that $\Omega \not< \Omega$. Another true contradiction. So, for Priest, the less-than relation on ordinals is thus both well-founded (since it is a well-ordering) and it fails to be well-founded (since $\Omega < \Omega$). Once we get this far, we might as well go on. There are ordinals $\Omega + 1$, 2Ω, Ω^2, Ω^Ω, etc. So, for example, $\Omega^\Omega < \Omega$ (and $\Omega^\Omega \not< \Omega$). Starting with Ω+1, we have gone 'beyond the limits of thought'. Depending on the logic and the ordinal arithmetic, we seem to have that if $\Omega \leq \alpha$ and $\Omega \leq \beta$, then $\alpha = \beta$. In particular, $\Omega = \Omega{+}1 = 2\Omega$.

In this dialetheic set theory, it seems, long transfinite recursions and inductions are straightforward. The constructible sets are defined by transfinite recursion over all ordinals—over Ω—although I presume that we are no longer interested in relative consistency proofs, since the whole system is inconsistent. We can then do transfinite recursions over L (now a set), which amounts to a recursion of length 2Ω.

Even if we manage to get past the conceptual problems, and become comfortable with true contradictions, there are serious technical obstacles to overcome. The ideal would be to develop a set theory whose only axiom is something like Frege's infamous Basic Law V or the above ordinal abstraction principle:

$$\begin{aligned}(\text{ORD-ABS}) \quad & \forall R_1 \forall R_2 [(\mathbf{WO}(R_1) \& \mathbf{WO}(R_2)) \\ & \rightarrow (\mathbf{ORD}(R_1) = \mathbf{ORD}(R_2) \equiv R_1 \simeq R_2)].\end{aligned}$$

The theory should be sufficiently powerful, obtaining results as deep and interesting as ordinary ZFC. Ideally, it should have an interesting account of long transfinite recursions and inductions, and other such oddities. And, of course, the dialetheic set theory should be non-trivial: it should not be the case that every sentence in the language is a theorem (as it would be if ex falso quodlibet were valid). In other words, the theory should be powerful, but not too powerful. Since, at present, no one knows how to do this, I will conclude this part of the treatment, noting the potential for an interesting research project.[9]

[9] There are ways to produce what appear to be non-trivial models of a dialetheic set theory, and so perhaps there is some promise in this direction (see Priest (2002, Part 3, Technical Appendix, §§3–4)). Other accounts of naïve set theory allow abstraction principles like the above (ORD–ABS) and Frege's Basic Law V, but block the contradictions by restricting the logic in other ways. See, e.g. Weir (1998) and Field (2003a, 2003b, 2004). Considerations of these here would take us too far afield. Such theories typically take on the more standard revenge issues, involving truth. Their solutions to the Burali–Forti paradox stand or fall with their resolutions of the semantic paradoxes.

14.7 Summing Up

In closing, let me briefly recapitulate the various options on the Burali–Forti paradox, at least as concerns set theory and ordinal theory, along with a brief summary of what is wrong with each option.

(A) First, one can refuse to quantify over, or otherwise talk about all ordinals, or all von Neumann ordinals at once. The idea is that any quantifier over ordinals is necessarily restricted to only some of them. Of course, one cannot consistently *say* that this is what one is doing. In general, it is kind of hard to say just what it is that one cannot say. Here, since one must say that one cannot talk about *all ordinals*, one must violate the restriction in order to state it. Perhaps one is reduced to a Wittgensteinian 'showing', and not saying—although I have no idea how one should do that here. Or to pick up another Wittgensteinian metaphor, perhaps there is a ladder to be kicked away. But damned if I know how to put that either. 'Whereof one cannot speak, thereof one must be silent' (*Tractatus* 7). Not much chance of that.

The Zermelo program is one manifestation of this option. Any particular statement in the language of set theory is understood as being about an unspecified but determinate model of set theory. The problem, again, is that the first option does not allow one to *describe* the Zermelo program itself. That is, we cannot say things like 'there is an indefinitely extensible hierarchy of models', since that is to speak, albeit indirectly, of all von Neumann ordinals. So let us move on.

(B1) One can allow unrestricted quantification over all ordinals, but fail to acknowledge a totality or even a plurality of all ordinals. We just do not talk about 'the ordinals' (oops—just did); there just is/are no such thing(s) as 'the ordinals' to talk about. Arguably, there is no 'it', no single thing that 'contains' all ordinals. But if the ordinals exist, then surely we can talk about *them*. Why can't we just use the phrase 'the ordinals', and thus say things about them?

(B2) One allows unrestricted quantification over all ordinals—perhaps via the Zermelo program—but refuses unrestricted second-order variables ranging over the ordinals. This is what I keep calling the thin straw. It blocks the Burali–Forti paradox by not allowing one to state that the ordinals are well-ordered—even though we know perfectly well that they are. Indeed, we can prove that membership is transitive on the von Neumann ordinals, and we know that membership is well-founded. It follows that the von Neumann ordinals are well-ordered, or at least it would follow if we could only state the theorem. Moreover, we can easily define relations on the ordinals that seemingly constitute well-orderings strictly longer than that of the ordinals. All we cannot do is to state that these relations are well-orderings. This option demands that we give up what seem to be perfectly sound and legitimate expressive resources.

(C) One allows unrestricted second-order quantification over all ordinals, and thus holds that the ordinals are well-ordered. This allows one to define well-orderings that are strictly longer than the ordinals, but one denies that these long well-orderings have order-types, in the original sense of the term. After all, we saw at the outset (§14.1 above) that the abstraction principle (ORD—ABS) is inconsistent: some well-orderings do not have order-types. Since the long well-orderings are indeed well-orderings, it is (or seems to be) legitimate to do transfinite recursions and inductions over them.

There does not seem to be any principled reason why one cannot introduce terms for these well-orderings, either as new singular terms, as plural terms, or as higher-order constants. Of course, these new entities—things, pluralities, whatever—are not ordinals as originally understood. Rather, they are 'higher-order' ordinals, 'proper' ordinals, or 'super-ordinals'.

There may be some hypocrisy here. We thought that ordinary set theory was supposed to cover the ordinals in a maximally general sense. If we left out some well-orderings, then the hierarchy was not as large as it was envisioned to be. Well, maybe we can live with this. More important, the Burali—Forti issue iterates. Can we talk about *all* of the super-ordinals? I suppose that one can turn to a version of option (A), (B1), or (B2) above at this level, and deny oneself some obvious expressive resources when trying to talk about super-ordinals. We might allow unrestricted second-order quantification, but not unrestricted third-order quantification. Failing that, we are led to super-duper ordinals, and onward. Clearly, this option is unstable.

(D) One denies that any ordinals exist. This, of course, is the nominalist route to safety. If our nominalist does not wish to do without set theory—the baby in the dirty bath water—she might go on to develop a modalized theory, talking about possible collections and possible well-orderings. To keep Burali—Forti under control, we saw that the expressive resources must be limited, much along the lines of the limitations of ordinary set theory. The problem, as I see it, is to motivate the restrictions without invoking possibilia or possible worlds, even as metaphors. We need to know what the metaphors are metaphors for.

(E) One allows unrestricted quantification and allows the associated order-types. So Ω is a single object, an ordinal in the sense originally intended. So are $\Omega+1$, Ω^{Ω}, and the like. In other words, there are ordinals that come later than all the ordinals. To follow this route, of course, one must get over the natural rejection of contradictions, or tinker with the logic in some other way. Priest (1987), plus his later work, attempts to pave the way for this key move. But even if we manage this conceptual *Gestalt* shift, there are serious technical problems which must be resolved before a satisfactory solution of the Burali—Forti paradox is at hand. We need a non-trivial, non-ad-hoc, paraconsistent set theory as powerful as ZFC.

As the reader may have guessed, I am not very comfortable with any of these options. Every one of them has difficulties that would demand its rejection if the other options

were not just as bad. For what it is worth, my own sympathies, for now, are that the best tradeoff lies with the Zermelo program, either straight or modalized (D), where we grasp the thin straw of not allowing unrestricted second-order quantification when describing the hierarchy of models of set theory (B2). But it hurts to admit that. Moreover, I have no argument for my preference, and it hurts to admit that as well.

Much of this chapter is an extension of some ideas originally sketched in Shapiro and Wright (2006). In many cases I no longer remember who originally came up with each of the ideas. Thanks also to Kevin Scharp for his helpful comments on an earlier version of this chapter, and to the Arché Research Centre at the University of St Andrews, for devoting a session to the project.

References

Benacerraf, P., and Putnam, H. (1983). *Philosophy of Mathematics*, 2nd edn., Cambridge, Cambridge University Press

Boolos, G. (1984). 'To be is to be a value of a variable (or to be some values of some variables)', *Journal of Philosophy* 81, 430–50; reprinted in Boolos (1998a), 54–72

——— (1985). 'Nominalist platonism', *Philosophical Review* 94, 327–44; reprinted in Boolos (1998a), 73–87

——— (1985a). 'Reading the *Begriffschrift*', *Mind* 94, 331–44; reprinted in Demopoulos (1995, 163–181); reprinted in Boolos (1998a), 155–70

——— (1998a). *Logic, Logic, and Logic*, Cambridge, Mass., Harvard University Press

——— (1998b). 'Reply to Charles Parsons' "Sets and classes" ', in G. Boolos, *Logic, Logic, and Logic*, Cambridge, Mass., Harvard University Press, 30–6

Dummett, M. (1991). *Frege: Philosophy of Mathematics*, Cambridge, Mass., Harvard University Press

Feferman, S. (1991). 'Reflections on incompleteness', *Journal of Symbolic Logic* 56, 1–49

Field, H. (2003a). 'A revenge-immune solution to the semantic paradoxes', *Journal of Philosophical Logic* 32, 139–77

——— (2003b). 'The semantic paradoxes and the paradoxes of vagueness'. In JC Beall (ed.), *Liars and Heaps: New Essays on Paradox*, Oxford, Oxford University Press, 262–311

——— (2004). 'The consistency of the naïve theory of properties', *The Philosophical Quarterly* 54, 78–104

Glanzberg, M. (2004). 'Quantification and realism', *Philosophy and Phenomenological Research* 69, 541–72

Hellman, G. (1989). *Mathematics without Numbers*, Oxford, Oxford University Press

——— (2002). 'Maximality vs. extendability: reflections on structuralism and set theory'. In D. Malament (ed.), *Reading Natural Philosophy*, La Salle, Ill., Open Court, 335–61

Lakatos, I. (1976). *Proofs and Refutations*, ed. J. Worrall and E. Zahar, Cambridge, Cambridge University Press

Menzel, Christopher (1986). 'On the iterative explanation of the paradoxes', *Philosophical Studies* 49, 37–61

Parsons, C. (1977). 'What is the iterative conception of set?'. In R. Butts and J. Hintikka, (eds.), *Logic, Foundations of Mathematics and Computability Theory*, Dordrecht, Holland, D. Reidel, 335–67; reprinted in Benacerraf and Putnam (1983), 503–29; and Parsons (1983), 268–97

_____ (1983). *Mathematics in Philosophy*, Ithaca, NY, Cornell University Press

Priest, G. (1987). *In Contradiction: A Study of the Transconsistent*, Dordrecht, Martinus Nijhoff Publishers

_____ (2002). *Beyond the Limits of Thought*, 2nd edn. Oxford, Oxford University Press

_____ (2006). *Doubt Truth to be a Liar*, Oxford, Oxford University Press

Schimmerling, Ernest (2001). 'The ABC's of mice', *Bulletin of Symbolic Logic* 7, 485–503

Shapiro, S. (1991). *Foundations without Foundationalism: A Case for Second-Order Logic*, Oxford, Oxford University Press

_____ (1993). 'Modality and ontology', *Mind* 102, 455–81

_____ (1999). 'Do not claim too much: second-order logic and first-order logic', *Philosophia Mathematica* (3) 7, 42–64

_____ (2003). 'All sets great and small: and I do mean ALL', *Philosophical Perspectives* 17, 467–90

Shapiro, S., and Wright, C. (2006). 'All things indefinitely extensible', In A. Rayo and G. Uzquiano (eds.), *Absolute Generality*, Oxford, Oxford University Press

Weir, A. (1998). 'Naïve set theory is innocent', *Mind 107*, 763–98

Zermelo, E. (1930). 'Über Grenzzahlen und Mengenbereiche: Neue Untersuchungen über die Grundlagen der Mengenlehre', *Fundamenta Mathematicae 16*, 29–47; translated as 'On boundary numbers and domains of sets: new investigations in the foundations of set theory'. In William Ewald (ed.), *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*, Vol. 2, Oxford, Oxford University Press, 1996, 1219–33

15

Revenge and Context

Keith Simmons

15.1 Direct and Second-order Revenge

It's hard enough to find a satisfying response to the paradoxes, but the phenomenon of revenge can make it seem impossible. In the simplest manifestation of revenge—call it *direct revenge*—the pathological sentence or expression re-emerges intact from the attempt to treat it. This is familiar in the case of the liar. Ignorant of my whereabouts, I write on the board:

(L) The sentence written on the board in room 101 is not true.

But once it's realized that (L) is written on the board in room 101, we can reason in the usual way to the conclusion that (1) is in some way pathological, perhaps gappy or ungrounded. But if a sentence is pathological, it isn't true. So the sentence written on the board in room 101 is not true. And this true conclusion says just what (L) says—so the treatment of (L) as pathological has led to the re-emergence of (L) as a truth.

Though it's rarely noted, paradoxes of denotation and versions of Russell's paradox are also subject to direct revenge. Suppose I write the following expressions on the board:

(A) pi
(B) six
(C) the sum of the numbers denoted by expressions on the board in room 102.

But I've unwittingly written these expressions in room 102. So we can reason as follows. Suppose that (C) denotes a number, say k. Then $k = \pi + 6 + k$, which is absurd. So (C) is pathological, and fails to denote. (A) and (B) denote π and 6 respectively, and (C) fails to denote, it follows that the sum of the numbers denoted by expressions on the board in room 102 is $\pi + 6$. But the definite description in the previous sentence that refers to $\pi + 6$ is composed of exactly the same words as (C) with the same meanings. So (C) re-emerges as an expression that successfully denotes a number.[1]

Or consider a version of Russell's paradox. Suppose I write on the board (in room 103, of course):

(D) moon of the Earth

(E) unit extension of a predicate on the board in room 103

(where a unit extension is an extension with exactly one member). We now reason: the extension of (D) is a unit extension. So the extension of (D) is a member of the extension of (E). But does (E) have an extension? If it does, the extension is either self-membered or it isn't. Suppose first that it is. Then the extension of (E) has two members, so it is not a unit extension—and so it is not a self-member. Suppose second that it is not a self-member. Then the extension of (E) has just one member, so it is a unit extension—and so it is a self-member. Either way, we have a contradiction, and we conclude that (E) fails to have an extension. Now if (E) does not have an extension, then in particular it does not have a unit extension. So the only unit extension of a predicate on the board in room 103 is the extension of (D). But in the previous sentence there is a predicate composed of the same words as (E) with the same meanings. And since this predicate has a well-determined extension, so does (E).

In all these cases, our reasoning to the conclusion that (L) or (C) or (E) is pathological is not the end of the matter. We reason past pathology; indeed it is the pathological nature of these expressions that provides for their rehabilitation. (L) is true because it is pathological and so not true, which is what it says it is. (C) denotes $\pi + 6$ since, given that (C) is pathological, the sum of the numbers denoted by expressions on the board is $\pi + 6$. And (E) has a well-determined extension, since, given that (E) is pathological, the only unit extension on the board in the extension of (D). But once rehabilitated, (L), (C), and (E) can exact their revenge: if (L) is true, then it isn't; if (C) denotes $\pi + 6$, then the sum of the numbers denoted by expressions on the board is $\pi + 6 + (\pi + 6)$, so that $\pi + 6 = \pi + 6 + (\pi + 6)$; and if (E) has an extension with just one member, then it has an extension with two members, the unit extension of (D) and the unit extension of (E).

[1] Even more tightly self-referential is the phrase discussed in Hilbert and Bernays (1939): 'the successor of the integer denoted by this phrase'. These paradoxes of denotation are related to Richard's paradox (Richard 1905), König's paradox (König 1905), and Berry's paradox (reported in Russell 1906 and 1908).

Direct revenge, then, makes life very difficult: we surely must conclude that these paradox-producing expressions are pathological in some way or other. But if we do, that seems only to encourage their immediate recovery and restore their power to produce paradox. It seems that we cannot call them pathological on pain of paradox! But if (L), (C), (E), and their ilk are not pathological, what are they?

Direct revenge is generated by the very sentences and expressions that we were trying to treat in the first place. But revenge can take another form—call it *second-order revenge*. Often a solution to paradox will introduce new, perhaps technical notions—for example, gaps (in truth, reference, or predicate-application), levels of a hierarchy, groundedness, determinate truth, stability, context. Second-order revenge takes these new notions, and constructs new paradoxes for old. Theories of truth, for example, face new challenges presented by sentences that say of themselves that they are false or gappy, or not true at any level of the hierarchy, or ungrounded, or not determinately true, or not stably true, or not true in any context.[2]

The connection between direct and second order-revenge is a delicate matter. The notions that generate direct revenge—truth, denotation, extension—are the initial targets of an attempt to solve the paradoxes. Those that generate second-order revenge appear to be more specialized semantic notions, ingredients of a semantic theory that deals with paradox. Yet theorists are likely to present these notions as themselves natural and intuitive—the solution should not be artificial, unconnected to our ordinary semantic intuitions. But the more natural these notions, the more they should be regarded as an initial target. For example, a gap theorist is likely to appeal to the naturalness of the notion of a truth-gap. And if truth-gaps are part of our ordinary repertoire, then so is the disjunctive notion of being *false or gappy*, along with the coextensive notion of being *not true*, on one natural reading of negation. Here, second-order revenge collapses into the first-order revenge generated by (L)—and then so much the worse for the gap theorist, if the theory cannot deal with even the initial target.

Where there is no such collapse, second-order revenge presents a distinct challenge to a semantic theory. Suppose the newly introduced concepts, though natural enough, are not part of our ordinary repertoire, and so are inappropriate initial targets. But since they do give rise to paradox, the theory is limited—even if it can deal with the initial targets, it cannot deal with these new ones. This is a significant failure: on pain of paradox, the semantic theory cannot accommodate natural enough semantic concepts. Second-order revenge seems to present an unpalatable choice, between contradiction on the one hand, and a significant expressive incompleteness on the other. Second-order revenge threatens to show that however successfully a theory deals with its initial targets, it cannot deal adequately with the general phenomenon of semantic paradox.

[2] See Herzberger (1980)–1 for a vivid demonstration of the problem posed by revenge Liar paradoxes. Herzberger (1970) discusses one kind of second-order revenge paradox–paradoxes of grounding.

15.2 Dealing with Direct Revenge

Direct revenge seems to put the semantic theorist in a bind. Any account of paradox will surely characterize (L) or (C) or (E) as pathological in some way or other. For example, (L) is characterized by Kripke as ungrounded,[3] in one treatment of his paradox of definability, Richard suggests that the analogue of (C) fails to denote,[4] and Martin and Maddy suggest that the extension of a Russell predicate, analogous to (E), falls into a membership gap, failing to belong to its own extension or anti-extension.[5] But direct revenge seems to show that the withholding of a truth-value to (L), of successful reference to (C), or of a determinate extension to (E), leads only to the reinstatement of a truth-value, a reference, or an extension, and the apparent return of paradox.

But I believe there is a natural way of treating direct revenge. Rather than regarding the reasoning as a threat, we can regard it as intuitive and valid reasoning that instructs us about our use of semantic concepts. There is a common pattern in all three direct revenge discourses. We start with the paradox-producer, the expression (L) or (C) or (E). We reach the conclusion that it is pathological. We then go on to *repeat* the very expression that caused the problem in the first place. But when we repeat the expression, when we use the very same words that compose (L) or (C) or (E), we find that the repeated expression *does* have a determinate truth-value, reference, or extension. In each case, we have two tokens of the same type—one pathological, one not. That is, we have two tokens of the same type but with different semantic status. The natural thought is that these tokens are produced in different context with different results. This is a thought that I've explored elsewhere.[6] Let me sketch the main ideas here.

It's a familiar idea that *context acts on content*—consider indexicals like 'I' and 'now'. But it is increasingly being recognized that this is not a one-way street. The reverse direction holds as well: *content acts on context*. Stalnaker writes:

> context constrains content in systematic ways. But also, the fact that a certain sentence is uttered, and a certain proposition expressed, may in turn constrain or alter the context . . . There is thus a two-way interaction between contexts of utterance and contents of utterances.[7]

At a given point in a discourse, the context will in part depend on what has been said before. For example, the context may change as new information is added to the discourse. Over the last twenty years or so, the kinematics of context-change has been studied by philosophers, semanticists, and linguists alike.[8]

[3] Kripke (1975). [4] Richard (1905). [5] Martin (circulated xerox); Maddy (1983).

[6] Simmons (1993) deals with truth, Simmons (1994) with denotation, and Simmons (2000) with extensions.

[7] Stalnaker (1975), in Stalnaker (1999), p. 66. [8] For a survey article, see Muskens *et al.*(1997).

According to Stalnaker the connection between context and available information is very tight indeed. Stalnaker writes:

> I propose to identify a context (at a particular point in a discourse) with the body of information that is presumed, at that point, to be common to the participants in the discourse.[9]

To put it another way, a context is to be represented by the shared presuppositions of the participants[10]—or the 'common ground', to use a phrase from Grice.[11] As new utterances are produced, and new information is made available, the context changes. For a specific example, consider the speech act of assertion: 'Any assertion changes the context by becoming an additional presupposition of subsequent conversation.'[12]

The shared presuppositions of conversants also figure in David Lewis's account of context-change. Lewis introduces the notion of a *conversational score*.[13] Following Stalnaker, Lewis identifies the set of shared presuppositions of the participants (at a given stage of a conversation) as one component of the conversational score. 'Presuppositions can be created or destroyed in the course of a conversation'[14]—and as the set of presuppositions changes, the conversational score changes. Of course, the notion of conversational score is a vivid way of capturing the notion of context. A change in the set of presuppositions is a change of context.

[9] Stalnaker (1988), in Stalnaker (1999), p. 98. This is a repeated theme in Stalnaker's writings; for example: '. . . a context should be represented by a body of information that is presumed to be available to the participants in the speech situation' (Stalnaker (1999), p. 6).

[10] Stalnaker takes presuppositions here to be *pragmatic presuppositions*: 'Presuppositions, on this account, are something like the background beliefs of the speaker propositions whose truth he takes for granted, or seems to take for granted, in making his statement'. (Stalnaker (1974), in Stalnaker (1999), p. 48)

[11] See Grice's 'Logic and Conversation' (The William James lectures), Part 1 of Grice (1989).

[12] Appendix to Stalnaker (1975), in Stalnaker (1999), p. 77. In a similar vein, Stalnaker writes: '. . . the essential effect of an assertion is to change the presuppositions of the participants in the conversation by adding the content of what is asserted to what is presupposed'. (Stalnaker (1978), in Stalnaker (1999), p. 86.)

[13] The analogy is with a baseball score. A baseball score for Lewis is composed of a set of seven numbers that indicate, for a given stage of the game, how many runs each team has, which half of which innings we're in, and the number of strikes, balls, and outs. Notice that correct play depends on the score—what is correct play after two strikes differs from what is correct play after three strikes. Similarly for conversations: the correctness of utterances—their truth, or their acceptability in some other respect—depends on the *conversational* score. Lewis continues: 'Not only aspects of acceptability of an uttered sentence may depend on score. So may other semantic properties that play a role in determining aspects of acceptability. For instance, the constituents of an uttered sentence—subsentences, names, predicates, etc.—may depend on the score for their intension or extension.' (Lewis (1979), in Lewis (1983), p. 238) This last remark of Lewis's is particularly relevant, since we are concerned with the subsentential terms 'true', 'denotes', and 'extension'.

[14] Lewis (1979) p. 233.

Another component of the conversational score, according to Lewis, is the *standard of precision* that is in force at a given stage of the discourse. Suppose I say 'France is hexagonal.' If you have just said 'Italy is boot-shaped', and got away with it, then my utterance is true enough. The standards of precision are sufficiently relaxed. But if you have just denied that Italy is boot-shaped, and carefully pointed out the differences, then my utterance is far from true enough—the standards of precision are too exacting. The acceptability of what I say here depends on the conversational score, on the context, which in turn depends on what has been said before. The extension of 'hexagonal' shifts with changes of context. Or, for another example, suppose I say 'The pavement is flat' under standards of flatness where the bumps in the pavement are too small to be relevant. Then what I say is true. But if the conversational score changes, and I say 'The pavement is flat' under raised standards of flatness, what I say will no longer be true. But '[t]hat does not alter the fact that it *was* true enough *in its original context*'.[15] Like the extension of 'hexagonal', the extension of 'flat' changes with the context.[16]

Let's return to direct revenge. Let $\mathcal{P}$ stand for the pathological token (L) or (C) or (E). It is natural to divide our direct revenge discourses into four segments: first, where I produce tokens on the board; second, where we reason to the conclusion that $\mathcal{P}$ is pathological and fails to denote; third, where we repeat the pathological expression; and fourth where we conclude that the repetition, call it $\mathcal{P}^*$, and hence $\mathcal{P}$, have a semantic value (a truth-value, a referent, or an extension).[17]

Now consider in particular the transition from the second to the third segment of the discourse. The culmination of the reasoning of the second segment is the proposition that $\mathcal{P}$ is pathological and fails to have a value. This is new information, and the proposition becomes one of our shared presuppositions, part of the common ground. So in the transition from the second segment to the third, there is a context change—a shift in the body of information that is presumed to be available.[18]

[15] Lewis (1979) p. 246.

[16] The notions of common ground and shared presuppositions also figure in Irene Heim's *file change semantics*, where a 'file' contains all the information that has been conveyed up to that point—and the file is continually updated as the discourse moves on. For Heim, 'the common ground of a context be identified with what I have been calling the "file" of that context' (Heim (1988), p. 286). Heim's notion of common ground is more fine-grained than Stalnaker's, since Heim's account is more sensitive to the subsentential structure of sentences. As with Lewis, the extra fine-grainedness of Heim's account is of relevance to us, since the present concern is with the subsentential terms 'true', 'denotes', and 'extension'. But clearly, the accounts of Stalnaker, Lewis, and Heim are broadly similar—they track context-change in terms of shifts in the shared presuppositions or common ground of the participants.

[17] The division of the revenge discourse into segments fits naturally into Grosz and Sidner's dynamic theory of discourse structure (see Grosz and Sidner (1986)), but I cannot pursue the details here.

[18] According to Heim's account, I will register this shift by updating the file card that stores information about C: I will now add the entries 'is pathological' and 'does not denote a number'. Grosz and Sidner's focusing structure distinguishes the salient objects, properties, and relations at each point

Let us say that the new contexts associated with the third and fourth segments are *reflective with respect to C*. In general, a context associated with a given point of a discourse is *reflective with respect to a given expression* if at that point it is part of the common ground that the expression is semantically pathological, and fails to have a value. So as we move from the second segment to the third, there is a context-change—a shift to a context that is reflective with respect to $\mathcal{P}$. This context-change is an essential ingredient of direct revenge.

How do these changes in context act on content? Let *t* be the semantic term ('true', 'denotes', or 'extension') that appears in $\mathcal{P}$ and $\mathcal{P}^*$. Let 'i' denote the initial context in which $\mathcal{P}$ is produced, and let 't_i' represent a use of a token of *t* that is coextensive with the occurrence of *t* in $\mathcal{P}$. This by itself is not to commit us to the claim that *t* is context-sensitive, that the extension of *t* may shift with context—for all that we've said so far, it may be that all uses of *t* are coextensive, whatever the context. When we reach the conclusion that $\mathcal{P}$ is pathological, we will have employed a *t*-schema—either a truth schema ('s' is true iff s) or a denotation schema ('a' denotes b iff a = b) or an extension-schema (a is in the extension of 'F' iff Fa). In each case, we will have employed the t_i-schema, that is, the schema that contains an occurrence of t coextensive with the occurrence of t in $\mathcal{P}$. This is how a contradiction is reached; for example, we reach a contradiction by assessing the sentence (L) by the true_i-schema, obtaining (L) is true_i iff (L) is not true_i.

In the subsequent reflective context, we produce $\mathcal{P}^*$, a repetition of $\mathcal{P}$. The occurrence of *t* in $\mathcal{P}$ is again to be represented by 't_i'. Consider for example the repetition C* of C. Recall our reasoning: since C fails to denote, 'it follows that the sum of the numbers denoted by expressions on the board in room 102 is $\pi + 6$'. But C's failure to denote is a failure to denote_i—C's pathology is the result of its assessment by the denotes_i-schema. So we arrive at the sum $\pi + 6$ as the referent of C* by way of C's failure to denote_i a number. It is because C fails to denote_i, while A and B succeed, that the sum is $\pi + 6$. So our reasoning is represented as follows: since C fails to denote_i, it follows that the sum of the numbers denoted_i by expressions on the board in room 102 is $\pi + 6$. The occurrence of 'denotes' in (C*) is represented by 'denotes_i': the occurrence of 'denotes' in C* inherits its extension from that of the occurrence of 'denotes' in C. Similarly, it is because (E) fails to have an extension_i while (D) succeeds that we are led to the conclusion that the only unit extension of an expression on the board in room 103 is that of (D)—and so the occurrence of 'extension' in (E*) is represented by 'extension_i'. And the occurrence of 'true' in (L*) ('The sentence written on the board in room 101 is not true') is represented by 'true_i', since we infer (L*) from (L)'s failure to have a truth-value when assessed by the true_i-schema.

of the discourse—and as we move from the second segment to the third, and on to the fourth, it will distinguish the pathologicality of C and its failure to denote.

So $\mathcal{P}^*$ is a repetition of $\mathcal{P}$ in a strong sense: it is composed of the same words with the same meanings *and* the same extensions. And yet in our discourse we provide a definite value for $\mathcal{P}^*$. But if $\mathcal{P}^*$ is assessed by the t_i-schema, then, just as with $\mathcal{P}$, no value is forthcoming, So $\mathcal{P}^*$ is evaluated by a different schema, a schema associated with the reflective context in which $\mathcal{P}^*$ is produced, a schema that provides for the evaluation of $\mathcal{P}^*$ in the light of $\mathcal{P}$'s pathology. If we are to respect the reasoning, we must discern a shift in the extension of *t*, and a shift in the schema by which $\mathcal{P}^*$ is evaluated. For example, (C*) does not denote$_i$, but it does denote—let us say it denotes$_r$, to mark the shift in the extension of 'denotes' when we assess (C*) reflectively, in the light of C's pathology.

What produces this shift in the extension of *t*? The change in context—specifically, the shift to a context which is reflective with respect to $\mathcal{P}$. At the third stage, the reflective character of the context had the effect of disengaging $\mathcal{P}$ from the i-schema. Now, at the fourth stage, it has the effect of bringing into play a new schema—the reflective r-schema. When we assess $\mathcal{P}^*$, and declare that it has a value, we assess it in a context where it is part of the common ground that $\mathcal{P}$ is pathological. The schema by which we assess $\mathcal{P}^*$ provides an assessment of $\mathcal{P}^*$ *in the light of* $\mathcal{P}$*'s pathologicality*. For example, here's what the instance of the r-schema looks like in the case of (L*):

$$(L^*) \text{ is true}_r \text{ iff } (L) \text{ is not true}_i.$$

Given the information that now forms part of the common ground—*that (L) is pathological and is not true$_i$*—the right-hand side of the biconditional holds, and it follows that (L*) is true$_r$. Similarly for (C*) and (E*). The difference between the assessments of $\mathcal{P}$ and $\mathcal{P}^*$ is explained this way: $\mathcal{P}$ is assessed by the unreflective t_i-schema, and $\mathcal{P}^*$ by the reflective t_r-schema. With the change in context, there is a change in the implicated schema. There is no intrinsic difference between $\mathcal{P}$ and $\mathcal{P}^*$—the difference lies in the schemas by which they are assessed.

Notice that $\mathcal{P}$ also has a value when assessed by the t_r-schema, the same value that $\mathcal{P}^*$ does. So, for example, (E) fails to have an extension and also has an extension. But there is no contradiction here: (E) fails to have an extension$_i$, but does have an extension$_r$. Compare Lewis's treatment of 'hexagonal' or 'flat'. Sometimes an utterance of 'France is hexagonal' (or 'The pavement is flat') is true, and sometimes it isn't. The extension of the predicates 'hexagonal' or 'flat' depends on the conversational score, in particular on the standards of precision that are in force. In a loosely analogous way, whether or not it is true to say that (E) has an extension will depend on the standard of assessment: do we apply the unreflective t_i-schema or the reflective t_r-schema?

The upshot is that *t* is a context-sensitive term that may shift its extension, depending on the schema of assessment that is in force in the given context. This in turn depends on the common ground, the information that is presumed to be available—in particular, information concerning the pathologicality of

denoting expressions. When $\mathcal{P}$ is first assessed, the information that it is pathological is not part of the common ground. The initial schema of assessment is unreflective with respect to $\mathcal{P}$. Once the information that $\mathcal{P}$ is pathological is incorporated into the common ground, we have a new standard: the subsequent schema of assessment is reflective with respect to $\mathcal{P}$. We have identified a contextual parameter—the *reflective status* of a context—to which the term *t* is sensitive.

If we do not attend to our ability to reason past pathology, the claim that *t* is a context-sensitive term will come as a surprise, and *reflective status* will not be an obvious contextual coordinate (unlike the familiar co-ordinates of speaker, time, and place, for example). But once we pay careful attention to discourses where we reason past pathology, it is natural and intuitive to conclude that *t* is indeed sensitive to the reflective status of a context. Cresswell once wrote:

> It seems to me impossible to lay down in advance what sort of thing is going to count [as a relevant feature of context] . . . The moral here seems to be that there is no way of specifying a finite list of contextual coordinates.[19]

Along with Cresswell, Lewis, Stalnaker, and others, we should be open to contextual co-ordinates beyond the familiar ones. If we recognize that content acts on context, that new information or new presuppositions can change the context, then we can identify contextual co-ordinates that we might otherwise miss. Reflective status is such a co-ordinate. When we conclude that $\mathcal{P}$ is pathological, a new presupposition is created, and the context changes. And the context-change is a change in reflective status. The key difference between context *i* and context *r* is the difference in reflective status: context *i* is not reflective with respect to $\mathcal{P}$, but *r* is. If we are after the extension of some token of the indexical 'I', we must determine who the speaker is in the given context. If we want to know whether $\mathcal{P}$, or any semantically pathological expression, is in the extension of a token of *t*, we must determine whether or not the given context is reflective with respect to that expression.

It is not initially obvious that 'true' or 'denotes' or 'extension' are context-sensitive, or that reflective status is a contextual coordinate along with speaker, time, and place. But if these claims provide the most plausible treatment of direct revenge, then surprise can give way to an improved understanding of how our semantic terms work. This is why we study paradoxes: we hope to learn from them.

[19] Cresswell, quoted in Lewis (1980), p. 30. One target of Cresswell's remark is Lewis (1970), and Lewis takes Cresswell's criticism to heart in Lewis (1980). In a somewhat similar vein, Kaplan writes: 'context *provides* whatever parameters are needed' (Kaplan (1989), p. 591), though Kaplan's remark is restricted to expressions that are 'directly referential'.

15.3 Second-order Revenge

Consider Kripke's theory of truth, and its central notion of groundedness. Though the object language $\mathcal{L}$ of Kripke's minimal fixed point is semantically closed with respect to its truth and falsity predicates, it is not with respect to 'grounded'. If we add the 'grounded' predicate to $\mathcal{L}$, a second-order revenge paradox is generated by, for example, 'This sentence is false or ungrounded' (contradiction follows whether we assume the sentence is true, false, or ungrounded). The escape route is an ascent to a metalanguage. Central terms of Kripke's theory, like 'grounded' and 'paradoxical', are not in the object language, but in a metalanguage in which the theory is expressed. Even if paradoxes involving truth and falsity are handled by Kripke's theory, paradoxes involving groundedness are not. The notion of groundedness is beyond the expressive capacity of $\mathcal{L}$.

This is typical of second-order revenge: the semantic theorist is forced to accept expressive incompleteness on pain of contradiction. We start with a target semantic notion—in Kripke's case, truth—and provide a theory of that notion which is not vulnerable to the associated paradoxes. In Kripke's case, we have a precise characterization of a language that can express consistently its own notion of truth. But nevertheless $\mathcal{L}$ is expressively incomplete—it cannot express the semantic notions introduced by the theory, such as groundedness. I have argued elsewhere that parallel remarks can be made about a variety of theories of truth and the semantic notions they introduce, whether stable truth, definite truth, determinate truth, or fuzzy truth, and so on.[20]

This failure of expressive completeness seems to compromise a theory's claim to resolve semantic paradox: a second-order revenge paradox is a semantic paradox beyond the scope of the given theory. How might the theorist respond? One response might go like this: these revenge paradoxes turn on technical notions, and the proper setting of the semantic paradoxes is ordinary language.[21] Terms like 'true' and 'denotes' are terms of ordinary language; terms like 'grounded' and 'stably true' are not. So, for example, if Kripke's minimal fixed point language $\mathcal{L}$ is a plausible model of English, then it's plausible to say that we have a solution to the liar in its natural setting. The problem with this response is that these introduced notions are supposed to be intuitive. We can readily grasp the thought that the evaluation ' "Snow is white" is true' is grounded in a sentence free of the truth predicate, while 'This sentence is true' is not; or the idea that the truth value of 'This sentence is false' is unstable, flip-flopping between truth and falsity (if it's true, then it's false, so then it's true, so then it's false . . .); or the claim that 'Harry is bald' is true' can be regarded as no

[20] See Simmons (1993), esp. chs. 3 and 4.

[21] Kripke suggests a response along these lines in Kripke (1975), pp. 79–80 and fn. 34.

more definitely true than 'Harry is bald'; and so on. Indeed, if these notions were not natural and intuitive, the theories would face the charge that they're artificial and unmotivated. So the objection remains: the theories cannot deal with semantic paradoxes generated by natural enough semantic notions.

A second response might go like this: why expect the theory to deal both with the original target concepts *and* with the theoretical concepts of the theory itself? The basic concepts of truth, denotation, and extension are to be treated one way, and the theoretical concepts another. For example, why not treat the revenge paradoxes that turn on groundedness or stable truth by a distinction between levels of language, and treat the language of the theory as a metalanguage for the target object language? The problem with this response is twofold. First, the family of revenge paradoxes, both direct and second-order, seems too close-knit to require distinct kinds of resolution. The sentences that generate second-order revenge ('This sentence is false or ungrounded', 'This sentence is not stably true', etc.) seem very like those that generate direct revenge, and the contradiction-producing reasoning looks very similar. The concepts may be different, but the structure of paradox remains the same. Second, whatever additional way out is offered for the introduced concepts, that too will face its own second-order revenge. If, for example, we appeal to a distinction between language levels, then we face the challenge posed by 'This sentence is not true at any level.' The problem of second-order revenge is just postponed.[22]

If expressive incompleteness signals a failure to deal with paradox, and if second-order revenge forces expressive incompleteness on any consistent theory, then perhaps inconsistency is the price we should pay. According to the dialetheist, there are true contradictions, and liar sentences, for example, are both true and false. In classical logic, of course, everything follows from a contradiction—and the dialetheist cannot allow that everything is true. So the contradictions associated

[22] In a series of recent papers (see e.g. Field (2003)), Field has offered a formally sophisticated treatment of the liar, which focuses on the revenge problem and achieves a remarkably high degree of semantic closure. But ultimately, it seems to me, it is subject to the familiar complaint: the price of consistency is expressive incompleteness. Prominent features of Field's theory are a non-classical conditional, a three-valued logic, and the notion of determinate truth. Field defines a 'determinately' operator that iterates into the transfinite, so that, it is claimed, for every liarlike sentence there is an operator that expresses its semantic status. But none of these operators captures the general notion of determinate truth, and that seems like a notion that we have the resources to express. (On this, see Priest (2005), pp. 44–6, and Beall (2005), pp. 23–4.) In a similar vein, Yablo distinguishes two questions:

Question 1: is the language able to characterize as defective every sentence that deserves to be so characterized? Question 2: are there intelligible semantic notions such that paradox is avoided only because these notions are not expressible in the language.

Yablo goes on to point out that a 'revenge-monger' may focus on Question 2, paying particular attention to the notion *having an ultimate value other than 1*, where the value 1 is to be understood as determinate truth.

with the paradoxes are quarantined by some suitable paraconsistent logic. Accept these quarantined contradictions and the paradoxes are tamed. The very notion of revenge seems misplaced now, for what worse could a purported revenge paradox produce than a contradiction? For the price of inconsistency we can buy expressive completeness. However, despite appearances, there are revenge paradoxes for the dialetheist. Since dialetheists focus mainly on truth, I shall focus here on revenge *liars*.

According to the dialetheist, some sentences relate just to the value T ('$2 + 2 = 4$'), some relate just to the value F ('$2 + 2 = 5$'), and some, like the liar sentences, relate both to T and to F.[23] (Some dialetheists, though not Priest, will also allow that there are sentences that relate to neither value. For simplicity, I'll set this form of dialetheism aside.) Let the *evaluation set* of a sentence be the set of values to which the dialetheist relates the sentence. So, for example, the evaluation set of '$2 + 2 = 4$' is the unit set {T}, the evaluation set of '$2 + 2 = 5$' is the unit set {F}, and the evaluation set of 'This sentence is false' is the set {T, F}. Now consider the sentence:

(X) The evaluation set of (X) is {F}.

Since X is a liar sentence, the dialetheist will say that it relates to both T and F. So we can infer:

(i) The evaluation set of (X) is {T, F}.

But since X is true, it follows that:

(ii) The evaluation set of (X) is {F}.

From (i) and (ii) we obtain that {T, F} = {F}, from which it follows that T = F.[24] But then, since everything is true or false for our dialetheist (there are no gaps), everything is true. This is unacceptable to the dialetheist—and we have a (second-order) revenge liar.[25]

How might the dialetheist respond? Perhaps along the following lines. The sentence (X) is related *just* to F (as it truly says of itself), and it's *also* related to both T and F. In terms of evaluation sets that is to say that the evaluation set of (X) is {F}, *and* that the evaluation set of (X) is {T, F}. This is inconsistent—but the dialetheist is not constrained by consistency when dealing with the liar.

[23] Priest presents dialetheism in relational terms in Priest (2002), ch. 4.6, and in Priest (2006), ch. 20.3.

[24] Perhaps the dialetheist might accept {T, F} = (F}, and also {T} = {F}, but deny T = F. This would be a denial of extensionality for at least impure set theory (given that T and F are not sets). But this move seems quite *ad hoc* and unmotivated. If we accept, as the dialetheist does, that T and F are distinct values, it is hard to make sense of the thought that the collection of both values is the same as the collection of just one of them.

[25] If gaps are admitted, we'll obtain that every sentence is true or gappy—also unacceptable to the dialetheist.

The trouble with this response is that, on dialetheist grounds, the evaluation set of a sentence will be unique. There will be just one set of values to which a sentence is related by the dialetheist account. If a dialetheist says that a sentence is related to T, then T is in the evaluation set of the sentence, whether or not the dialetheist also says that the sentence is related only to F. So, in particular, if the dialetheist says that (X) is related just to F, that is not to say that {F} is the evaluation set of (X)—it will depend on what *else* the dialetheist relates (X) to. Consider the likely dialetheist response to the sentence:

($L^{\dagger}$) This sentence is false only.

This liar sentence is true and false according to the dialetheist, and since it's true, it's false only. But the dialetheist will say that ($L^{\dagger}$) is false only, *and* both true and false. In relational terms, one can put it this way: ($L^{\dagger}$) is related just to F, and is also related to both T and F. Being false only does not rule out the additional information that ($L^{\dagger}$) is true too. In general, for the dialetheist, being false does not preclude being true, and neither does being false *only* preclude being true. When we list all the values to which ($L^{\dagger}$) is related by the dialetheist, there will be a single, determinate list, namely, the list T, F. Similarly for (X). The evaluation set for any liar sentence will be the unique set {T, F} because the dialetheist will relate any liar sentence to both values, whether or not she will also allow that it is false only (or true only) as well. So the problem posed by (X) remains: (X) has a unique evaluation set, with the consequence that T = F.

Notice also that dialetheists do not suggest that the semantic status of liar sentences is in any way unstable or ambiguous. It is no part of the dialetheist account that the truth-value of a liar sentence can somehow shift, say, from *true and false* to *false only*. According to some theories of truth—such as the revision theory of truth or contextual theories—we should pay close attention to shifts in our evaluations of liar sentences. But dialetheists reject these theories. According to the dialetheist, paradoxical sentences are supposed to receive a single, stable evaluation—they're true and false. An evaluation that was somehow ineffable or inexpressible would be quite against the spirit of dialetheism: part of the motivation for the view is to avoid counterintuitive restrictions on expressibility. For the dialetheist, the evaluation set of a Liar sentence does not change, and neither is it inexpressible.

Why does (X) pose a problem for the dialetheist? The case of ($L^{\dagger}$) provides a clue, I think. There, the dialetheist's response turns on a certain kind of open-endedness to the evaluation of ($L^{\dagger}$). If the value *false only* were to close off even the inconsistent addition of the value *true* (or vice versa), then ($L^{\dagger}$) would generate a revenge paradox. But the addition of *true* to *false only* (or vice versa) is acceptable by dialetheist lights: neither *false only* nor *true and false* counts by itself as a complete evaluation of ($L^{\dagger}$)—neither tells the whole story. But the notion of an *evaluation set* forces the dialetheist to tell the whole story about a sentence, in the sense that the members of the evaluation set will

exhaust the values to which the dialetheist relates the sentence. Since the dialetheist is committed to the truth of (X), (X) truly identifies its own evaluation set—that is, whatever values are in the identified set are the values to which the dialetheist relates (X). Given the uniqueness of the evaluation set, we obtain the result that {F} = {T, F}, and the consequence that T = F.

15.4 Dealing with Second-order Revenge

Let's return to the contextual view defended in 15.2. There I explained the difference between the assessments of $\mathcal{P}$ and $\mathcal{P}^*$ this way: $\mathcal{P}$ is assessed by the unreflective t_i-schema, and $\mathcal{P}^*$ by the reflective t_r-schema. There is a difference between the extension of the terms t_i and t_r—for one thing, $\mathcal{P}$ and $\mathcal{P}^*$ are in the extension of t_r, but not in the extension of t_i. What else is excluded from the extension of t_i? And what is the relation between the extensions of t_i and t_r?

A possible response here is a Tarskian one: t_i and t_r are associated with distinct levels of language.[26] The predicate t_r is associated with $\mathcal{P}$'s unreflective context of utterance; the predicate $\mathcal{P}^*$ is the more comprehensive denotation predicate of a semantically richer language associated with a context reflective with respect to $\mathcal{P}$. On such a hierarchical account, the extension of t_i is properly contained in the extension of t_r.

There are a number of prima-facie worries about such a hierarchical approach. For one, the stratification of English into a hierarchy of languages seems artificial. For another, we can ask how levels can be assigned to occurrences of *t* in a systematic way. But let me focus here on the problem of revenge for the hierarchical approach. Second-order revenge helps itself to the talk of levels and constructs new paradoxes generated by sentences and phrases such as (a) 'This sentence is not true at any level', or (b) 'the least ordinal not denoted by an expression at any level', or (c) 'non-self-membered extension of a predicate at some level'. Revenge here seems to force a choice: disallow unrestricted talk of all levels (expressive incompleteness) or ascend to a theoretical metalanguage which absorbs talk of all levels. The newly introduced semantic notions—*true at some level, denotes at some level, has an extension at some level*—apply more widely that any occurrence of *t*, which on the hierarchical account

[26] Tarski would not endorse this 'Tarskian' response to paradox in the setting of natural language. In Tarski (1933)/(1986), Tarski turns away from natural languages, and investigates only formal, regimented languages. The contextual accounts offered in Burge (1979) and Glanzberg (2001) are hierarchical in the sense that the extension of 'true' is tied to level, and the higher the level, the broader the extension of 'true'. In the singularity account that I'll be sketching below, the extension of 'true' is not tied to level in this way. In particular, the extension of t_r is not broader (or narrower) than t_i—neither extension includes the other.

must always occur at some particular level. For example, the predicate 'denotes at some level' applies to every denoting phrase at every level of the hierarchy—and no occurrence of the context-sensitive term 'denotes', appearing at some level of the hierarchy, applies so comprehensively. The pattern then is the familiar one: revenge is avoided either by admitting expressive incompleteness or by ascent to a metalanguage.

I want to suggest an alternative contextual account. Here I will provide a brief sketch, and then go on to consider revenge.[27] This alternative is in a strong sense anti-hierarchical: there is no stratification of the semantic term *t*. The leading idea is that *t* applies almost everywhere, except for certain singular points, or *singularities*.[28] More precisely, it is *occurrences* of *t* that have singularities. For example, $\mathcal{P}$ is a singularity of the occurrence of *t* in $\mathcal{P}$ ((L) is a singularity of the occurrence of 'true' in (L), (C) a singularity of 'denotes' in (C), and (E) a singularity of 'extension' in (E)). In general, suppose we are given a context α and a phrase or sentence σ containing the term *t*. If σ cannot be given t_α conditions, if it cannot be evaluated by the t_α-schema, then σ is a *singularity* of t_α. And if σ is a singularity of t_α, and α is σ's context of utterance, then σ is *pathological*. So, for example, (C) is a singularity of 'denotes$_i$', and it is pathological too, since the subscript stands for (C)'s context of utterance. (C*) is also a singularity of 'denotes$_i$', but (C*) is *not* pathological, since its context of utterance is the reflective context r, and (C*) does denote$_r$. Similarly with (L) and (L*), and (E) and (E*).

It's the job of the singularity theory to provide a systematic way of identifying the singularities of a given occurrence of *t*. Notice that something is missing if we represent, say, the token (C) as an ordered pair <type(C),i>, where the first element is the type of (C), and the second indicates the appropriate representation of 'denotes' in (C) (viz. 'denotes$_i$'). This representation does not distinguish (C) from (C*), yet the former denoting expression is pathological and the latter isn't. There is something more to consider: the schema by which (C) is given denotation conditions.

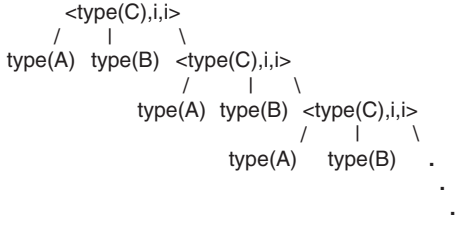
In the second segment of the revenge reasoning, (C) is evaluated by the i-schema; in the fourth segment, (C*) is evaluated by the r-schema, a schema that is reflective with respect to (C). So we capture the discourse more perspicuously if we represent (C) by the ordered *triple* <type(C),i,i>, where the third element indicates that the schema by which (C) is assessed is the i-schema, and (C*) by the triple <type(C),i,r>, indicating

[27] For full accounts, see Simmons (1993), (1994), (2000).

[28] In a tantalizing passage, Gödel writes: 'It might even turn out that it is possible to assume every concept to be significant everywhere except for certain 'singular points' or 'limiting points', so that the paradoxes would appear as something analogous to dividing by zero. Such a system would be most satisfying in the following respect: our logical intuitions would then remain correct up to certain minor corrections, i.e. they could then be considered to give an essentially correct, only somewhat "blurred", picture of the real state of affairs' (Gödel (1944), in Schilpp (1944), p. 228).

that (C*) is assessed by the r-schema. In the course of the revenge reasoning, (C) is evaluated by the i-schema, and it is this evaluation that leads to the conclusion that (C) is pathological; and (C*) is evaluated by the r-schema, and it is this evaluation that leads to the conclusion that (C*) has a determinate denotation. So if we are after an analysis of the revenge discourse, the representation <type(C),i,i> of (C) is privileged, and <type(C,i,r> is likewise a privileged representation of (C*). Call these the *primary representations* of C and C*.[29] In general, a primary representation of a sentence or phrase σ will represent σ as evaluated by the α-schema, where α is σ's context of utterance. We are interested in identifying semantic pathology, and if a token is pathological, it will lack a value if assessed by its associated α-schema. If pathology is there to be found, primary representations will help to reveal it.

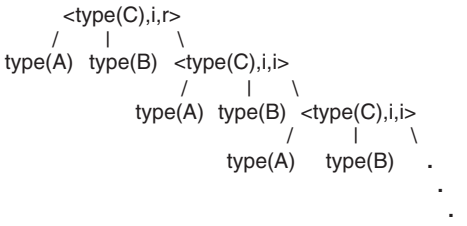
We will characterize the notions of *pathology* and *singularity* via a certain kind of tree. Consider, for example (C). (C) makes reference to denoting phrases, and to determine C's denotation we must first determine what these phrases denote—denote$_i$, that is, because that occurrence of 'denote' in (C) is represented by 'denotes$_i$'. So the appropriate schema by which to assess the phrases to which (C) refers is the i-schema. All this is captured by the *primary tree* for (C). The top node of the tree is the primary representation of (C), the triple <type(C),i,i> . This is the node at the top of the tree. At the second tier are the phrases to which (C) makes reference, namely (A), (B), and (C), where (C) is represented as evaluated by the i-schema (and for simplicity (A) and (B) are represented simply by their types, since they contain no context-sensitive terms). So the primary representation of (C) appears again at the second tier, and this in turn generates a third tier of nodes. And so on, indefinitely. The primary tree for (C) is unfounded:



We now say that (C) is pathological because *its primary representation repeats on its primary tree*. The unfounded tree shows that (C) cannot be assessed by the i-schema—and so we can also say that (C) is a *singularity* of 'true$_i$'.

[29] A *secondary representation* of a token σ containing *t* indicates the assessment of σ by a *t*-schema other than that associated with σ's context of utterance. As we'll have occasion to notice later, one secondary representation of (C) (<type(C), i, r>) is the prmary representation of C*—and one secondary representation of C* (<type(C), i, i>) is the primary representation of C.

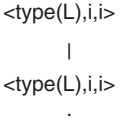
Contrast the primary tree for (C*). The primary representation of (C*) is <type(C),i,r> , and the primary tree for (C*) is:



The phrases at the second tier are represented as evaluated by the i-schema, since the occurrence of 'denotes' in (C*) is represented by 'denotes$_i$'. (More formally, the second member of the triple at the top of the tree is the third member of any ordered triple at the second tier.) At the second tier, then, we have the primary representation of (C), which then repeats at all subsequent tiers. But the representation of (C*) does not repeat, indicating that, unlike (C), (C*) is *not* pathological, and *not* a singularity of 'true$_r$'. Intuitively, (C*) stands above the circle in which (C) is caught. (C)'s pathology is not the end of the matter—we can reason past pathology. When we determine a value for (C*), we will need to determine the denotation$_i$ of the phrases to which is refers. Since (C) does not denote$_i$—as the tree indicates—we determine a value for (C*) in terms of (A) and (B) only.[30]

Notice that if we evaluate (C) by the r-schema—that is, evaluate it from a context that is reflective with respect to (C)—we will find that it does denote, just like (C*). This is shown by a *secondary tree* for (C). The triple <type(C),i,r> is a secondary representation of (C), since the third element is not (C)'s context of utterance. And the secondary tree for (C) is identical to the primary tree for (C*), indicating that (C), like (C*), does denote$_r$.

[30] The primary tree for (L) is an infinite single-branched tree:



The primary representation of (L) repeats on this tree, indicating that (L) is pathological and a singularity of 'true$_i$'. The primary tree for (L*) has <type(L),i,r> for its top node, and then every subsequent node of this single-branched tree will be the primary representation of (L). So while (L) is pathological, (L*) is not, since its primary representation does not repeat on its primary tree—and neither is (L*) a singularity of 'true$_r$'.

The primary tree for (E) is:

Now suppose that we evaluate (C) from some context other than i or r, where it is *not* part of the common ground that (C) is pathological. Then the corresponding secondary representation of (C) is <type(C), i, n> , where n is the neutral context. The secondary tree will be just like the primary tree of (C*), with 'n' replacing 'r'. And since this secondary representation does not repeat, the tree indicates that C does denote_n. This neutral context is treated as if it were reflective with respect to (C). We treat (C) as a denoting phrase—a phrase denoting $\pi + 6$—if we possibly can. We cannot allow that (C) denotes_i, but we can allow that it denotes_r and denotes_n. Restrictions on occurrences of 'denotes' are kept to a minimum.

The results delivered by the primary and secondary trees reflect the basic intuition behind the singularity account. Suppose you say ' "The square of 1" denotes 1.' Here, your use of 'denotes' is quite unproblematic. Should the pathological token (C) be excluded from its extension? The singularity account says no—because there is no need to exclude it. We have seen that (C) denotes_r $\pi + 6$ because the sum of the numbers denoted_i by expressions on the board *is indeed* $\pi + 6$. And for the same reason, (C) can be counted as a denoting expression in your neutral context of utterance. We have no reason to suppose that (C) must be evaluated from your context of utterance by the contradiction-producing i-schema; instead, your use of 'denotes' is treated as reflective with respect to C. This seems plausible: in general, speakers do not usually aim to produce pathological utterances, or utterances implicated in paradox.

Further, the singularity account respects a basic intuition about predicates. Intuitively, we take a predicate to pick out everything with the property that the predicate denotes. The more restrictions we place on occurrences of the semantic term *t*, the more we are at odds with this intuition. We do expect any solution to a genuine paradox to require some revision of our intuitions. But the more a solution conflicts with our intuitions, the less plausible that solution will be.[31]

This tree shows that (E) is pathological and a singularity of 'extension_i'. It's easy to see that the primary representation of (E*) does not repeat on its primary tree, and (E*) is not a singularity of 'extension_r'.

[31] For example, the Tarskian stratification involves massive restrictions on occurrences of *t*. On a standard Tarskian line, the referring expression 'the only even prime', for example, is of level 0; your unproblematic denoting phrase ('the number denoted by "the only even prime" ') is of level 1, and so

An advantage of the singularity account is, I think, that it provides a unified account of the semantic paradoxes, according to which the scope of each occurrence of *t* (whether 'true', 'denotes', or 'extension') is as close to global as it can be. And the corresponding *t*-schema is as close to unrestricted as it can be. Once we have a formal theory that identifies singularities of given occurrences of *t*, we can offer minimally restricted *t*-schemas. For example, here is the the true_i-schema:

If 's' is not a singularity of 'true_i', then 's' is true_i iff s.

Similarly, we can provide close-to-unrestricted schemas for 'denotes_i' and 'extension_i'.

This is a brief sketch of the singularity account, but perhaps it is enough for present purposes.[32] Revenge paradoxes for this contextual account may seem to be generated by the sentence 'This sentence is not true in any context', or the phrases 'the least ordinal not denoted in any context' and 'non-self-membered extension of a predicate in some context'. To fix ideas, let $\mathfrak{L}$ be the language that the singularity theory is a theory of. For simplicity, $\mathfrak{L}$ is taken to be English without any context-sensitive terms, plus 'true', or 'denotes', or 'extension'. Let $\mathfrak{T}$ be the language in which the singularity theory is expressed. To anticipate a little, we'll pay close attention to the relation between the language $\mathfrak{L}$ and the language $\mathfrak{T}$. We'll see that the crux of the matter is this: $\mathfrak{T}$ is *not* a Tarskian metalanguage for $\mathfrak{L}$.

For ease of exposition, we'll work with just one case, the case of truth—but bear in mind that what we say about truth carries over directly to denotation and extension. If we restrict ourselves to the language $\mathfrak{L}$, any occurrence of 'true' is an occurrence of a context-sensitive term. But in the language $\mathfrak{T}$, we freely quantify over contexts, and we explicitly attach contextual subscripts to uses of 'true' in $\mathfrak{L}$. For example, I can say that the sentence (L) does not denote_i and that it does denote_r. And I can go on to say that (L) is true in some but not all contexts. In the language $\mathfrak{T}$, then, we will find the predicate 'true-in-$\mathfrak{L}$', where this is understood as the short form of 'sentence of $\mathfrak{L}$ that is true_α for some context α'. This predicate constant may be regarded as the truth predicate for $\mathfrak{L}$.

Now we can observe that no occurrence of the context-sensitive predicate 'true' of $\mathfrak{L}$ is coextensive with 'true-in-$\mathfrak{L}$'. We can establish this in two steps. First, every occurrence of 'true' will have singularities—for example, we can add to the innocent statement '$2 + 2 = 4$' is 'true' the paradox-producing 'but this very conjunct isn't'.

on, through the levels. Your use of 'denotes' in an utterance of level 1 has in its extension all referring expressions of level 0, *and no others*. So all sentences of level 1 and beyond are excluded from the extension of such a use of 'denotes'. Gödel remarks of Russell's type theory that '. . . each concept is significant only . . . for an infinitely small portion of all objects' (Gödel (1944), p. 149). A similar complaint can be made about a standard Tarskian account of *t*: for example, an ordinary use of 'denotes' will apply to only a fraction of all the denoting expressions.

[32] Detailed accounts of the singularity theory can be found in Simmons (1993), (1994), and (2000).

Second, notice that the pathological conjunct just produced will be true in a suitably reflective context, and so true in some context. So our conjunct is a singularity of 'true' in our 'innocent' statement, but is in the scope of 'true-in-$\mathfrak{L}$'. In this respect, the extension of 'true-in-$\mathfrak{L}$' will be broader than that of any occurrence of 'true'. So the question arises: isn't $\mathfrak{T}$ a Tarskian metalanguage for $\mathfrak{L}$?

The answer is no. The predicate 'true-in-$\mathfrak{L}$' is the truth predicate for $\mathfrak{L}$ in the sense that it applies to exactly the truths of $\mathfrak{L}$. The scope of 'true-in-$\mathfrak{L}$' is restricted to the expressions of $\mathfrak{L}$. In contrast, a given occurrence of 'true' applies to all truths except its singularities. It applies to any true sentence of any language, *as long as the sentence is not identified as a singularity*. In particular, the scope of an occurrence of 'true' extends to truths expressed in the language $\mathfrak{T}$, including those that cannot be expressed in the language $\mathfrak{L}$—for example, those sentences of $\mathfrak{T}$ containing the predicate 'true-in-$\mathfrak{L}$'. (For an example related to revenge, consider 'This sentence is not true-in-$\mathfrak{L}$'. This is a true sentence of $\mathfrak{T}$, since it is not a true sentence of $\mathfrak{L}$.) So in this respect, any occurrence of 'true' is more comprehensive than the predicate 'true-in-$\mathfrak{L}$', since the scope of 'true-in-$\mathfrak{L}$' is limited to the sentences expressed in $\mathfrak{L}$.

Further, $\mathfrak{T}$ is in certain respects expressively weaker than $\mathfrak{L}$. $\mathfrak{T}$ is a 'scientific' language in which we describe the semantics and pragmatics of a context-sensitive term. In $\mathfrak{T}$ we take context-sensitive language to be the object of our study, and stand back from the contexts and the context-sensitive term that we are describing. We formally define notions like *primary tree* and *singularity*. In scientifically describing the behaviour of the context-sensitive term 'true', we do not use context-sensitive terms. There are no context-sensitive terms in $\mathfrak{T}$. When we present the singularity theory, we take up an abstract, theoretical point of view. $\mathfrak{T}$ is not a language that contains a term tied to context—it is *about* a language that contains a term tied to context.

But if $\mathfrak{T}$ is a classical scientific language free of context-sensitive terms, vagueness, ambiguity, and so on, it is provably subject to expressive limitations. In particular, $\mathfrak{T}$ cannot contain *its* own truth predicate, 'true-in-$\mathfrak{T}$'. For if it did, we could form the sentence 'This sentence is not true-in-$\mathfrak{T}$' within $\mathfrak{T}$, and derive a contradiction. The truth predicate for $\mathfrak{T}$ must be contained in a metalanguage for $\mathfrak{T}$, a language essentially richer than $\mathfrak{T}$. And this metalanguage will contain truths that cannot be expressed in $\mathfrak{T}$. Now none of these truths will be identified as singularities of an occurrence of the ordinary predicate 'true' of $\mathfrak{L}$. So the scope of 'true' extends beyond the reach of $\mathfrak{T}$. In this regard, $\mathfrak{L}$ is essentially richer than $\mathfrak{T}$. Clearly $\mathfrak{T}$ is no Tarskian metalanguage for $\mathfrak{L}$. Notice that the metalanguage for $\mathfrak{T}$ is subject to the same kind of expressive limitation, and we are led to a hierarchy of languages with $\mathfrak{T}$ at its base. Each metalanguage in the hierarchy contains truths that cannot be expressed at any lower level. But none of these expressions is a singularity of a given occurrence of 'true', and so they are all within its scope. In this respect, $\mathfrak{L}$ is essentially richer than any language in the hierarchy.

So $\mathfrak{T}$ is not a metalanguage for $\mathfrak{L}$. But perhaps this should not come as any surprise. Consider any semantic theory of a context-sensitive term. The language of the theory will be 'austere', free of indexical terms, 'scientific'. The language in which we give an adequate semantics for 'I' or 'now' will not itself contain the indexical 'I' or 'now'. There is no requirement that a theory of 'I' or 'now' provide context-sensitive means for referring to myself or the present time. So we should not expect an utterance of 'I am hungry' or 'The meeting starts now' to be translatable into the language of the theory. In general, we should not expect that there will be a way of translating the context-sensitive term into a term or phrase of the language of the theory. Now a Tarskian metalanguage 'must contain the object language as a part', or at least it must be the case that 'the object-language can be translated into the metalanguage'.[33] So in general a language in which we state the theory of a context-sensitive term will not be a Tarskian metalanguage.[34] In particular, the singularity theory of 'true'—and, similarly, of 'denotes' and 'extension'—is not couched in a metalanguage for the object language $\mathfrak{L}$.

Perhaps it's worth stressing that the singularity approach is not hierarchical: the terms 'true', 'denotes', and 'extension' are not stratified into a series of increasingly comprehensive predicates. Instead we have a single, context-sensitive term, and each occurrence of the term has singularities that other occurrences do not have. No occurrence is more (or less) comprehensive than another; each occurrence is minimally restricted. This feature of the singularity account shouldn't be obscured by the hierarchy generated from $\mathfrak{T}$. This hierarchy is generated from a classical scientific language, and a parallel hierarchy could be generated from, say, the language of arithmetic or chemistry, or any suitably regimented language. In each case we will obtain an infinite series of truth predicates—one series starting with the predicate 'true in $\mathfrak{T}$', another with the predicate 'denoting expression in the language of chemistry', and so on. But these series are composed of predicate constants of the form 'true in L', limited to some suitable scientific language or metalanguage L. But our interest lies in the English predicate 'true'—and according to the singularity solution, this is a context-sensitive predicate that applies to true sentences at every level of all of these hierarchies.

There are of course legitimate questions about these hierarchies. What is their extent? Can we quantify over all their levels? Must we resort to schematic generalizations? If we were offering a Tarskian account of revenge, these questions would be critical. But we are offering a different kind of contextual account, and so these questions are less urgent (though no less interesting). They are questions that do not bear directly on the singularity theory, because the theory does not stratify 'true', 'denotes', or

[33] Tarski (1944), in Blackburn and Simmons (1999), p. 126.

[34] This point in stressed by Kaplan, in Kaplan (1997), p. 7.

'extension'. These semantic terms of the object language apply to the language of the theory, and beyond.[35]

References

Beall, JC (ed.) (2005). *Deflationism and Paradox*, Oxford University Press

Blackburn, Simon, and Simmons, Keith (1999). *Truth*, in the series *Oxford Readings in Philosophy*, Oxford University Press, Oxford

Burge, Tyler (1979). 'Semantical paradox', *Journal of Philosophy* 76: 169–98; reprinted with a postscript in Martin (ed.), (1984), 83–117

Glanzberg, Michael (2001). 'The liar in context', *Philosophical Studies* 103: 217–51

Gödel, Kurt (1944). 'Russell's mathematical logic'. In Schilpp (1944)

Grice, H. P. (1989). *Studies in the Way of Words*, Harvard University Press

Grosz, B., and Sidner, C. (1986). 'Attention, intention, and the structure of discourse', *Computational Linguistics* 12: 175–204

Heim, Irene (1988). *The Semantics of Definite and Indefinite Noun Phrases*, Garland Publishing, New York and London

Herzberger, Hans (1970). 'Paradoxes of grounding in semantics', *Journal of Philosophy* 145–67

——— (1980-1). 'New paradoxes for old', *Proceedings of the Aristotelian Society*

Hilbert, D., and Bernays, P. (1934, 1939). *Grundlagen der Arithmetik*, vol.1 Berlin: Springer (vol. 2, 1939)

Kaplan, David (1997). 'What is meaning?: explorations in the theory of meaning as use' (unpublished draft)

König, Julius (1905). 'On the foundations of set theory and the continuum problem'. In van Heijenoort (1967), pp. 145–9

Kripke, Saul (1975). 'Outline of a theory of truth', *Journal of Philosophy* 72: p. 690–716. Reprinted in Martin (ed.) (1984), pp. 53–81

Lewis, David (1970). 'General semantics', *Synthese* 22: 18–67

——— (1979). 'Scorekeeping in a language game', *Journal of Philosophical Logic* 8: 339–59; reprinted in Lewis (1983), pp. 233–49

——— (1980). 'Index, context, and content'. In S. Kanger and S. Ohman (eds.), *Philosophy and Grammar*, Dordrecht, Reidel; reprinted in Lewis (1998), pp. 21–44

——— (1983). *Philosophical Papers*, vol.1, Oxford University Press

——— (1998). *Papers in Philosophical Logic*, Cambridge University Press

Maddy, Penelope (1983). 'Proper classes', *Journal of Symbolic Logic* 48: 113–39

Martin, D. A. 'Sets vs. classes', circulated xerox

Martin, Robert L. (ed.) (1984). *Recent Essays on Truth and the Liar Paradox*, Oxford University Press

Muskens, R., van Benthem, J., and Visser, A. (1997). 'Dynamics'. In J. van Benthem and A. ter Meulen, (eds.), *Handbook of Logic and Language*, The MIT Press/North-Holland, pp. 587–648

[35] My thanks to Jamin Asay and Thomas Hofweber for their helpful comments on an earlier draft of this chapter.

Priest, Graham (2002). 'Paraconsistent logic'. In vol. 6, D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic* 2nd edn. Kluwer Academic Publishers, Dordrecht, pp. 287–393

_____ (2005). 'Spiking the Field-artillery' in Beall (2005)

_____ (2006). *In Contradiction*. 2nd edn., Oxford University Press

Richard, Jules (1905). 'Les principes des mathématiques et le problème des ensembles'. In *Revue générale des sciences pures et appliquées* 16, p. 541. Also in *Acta Mathematica* 30 (1906), pp. 295–6 (English translation in van Heijenoort (1967) pp. 143–4)

Russell, Bertrand (1906). 'Les paradoxes de la logique', *Revue du metaphysique et de morale* 14: 627–50

_____ (1908). 'Mathematical logic as based on the theory of types', *American Journal of Mathematics* 30: 222–262; reprinted in van Heijenoort (1967) (ed), pp. 150–82

Schilpp, P. A. (ed.) (1944). *The Philosophy of Bertrand Russell*, Open Court, pp. 123– 53

Simmons, Keith (1993). *Universality and the Liar: An Essay on Truth and the Diagonal Argument*, Cambridge University Press, Cambridge

_____ (1994). 'Paradoxes of denotation', *Philosophical Studies* 76: 71–104

_____ (2000). 'Sets, classes and extensions: a singularity approach to Russell's paradox', *Philosophical Studies* 100: 109–49

Stalnaker, Robert (1974). 'Pragmatic presuppositions'. In Milton K. Munitz and Peter Unger (eds.), *Semantics and Philosophy*, New York: New York University Press, pp. 197–213; reprinted in Stalnaker (1999), pp. 47–62

_____ (1975). 'Indicative conditionals', *Philosophia* 5: 269–86. Reprinted in Stalnaker (1999), pp. 63–77

_____ (1978). 'Assertion'. In Cole (ed.), *Syntax and Semantics, vol.9: Pragmatics*, New York: New York University Press, pp.197–213; reprinted in Stalnaker (1999), pp. 78–95.

_____ (1988). 'On the representation of context', *Journal of Logic, Language, and Information*, 7: 3–19

_____ (1999). *Context and Content: Essays on Intentionality in Speech and in Thought*, Oxford University Press.

Tarski, Alfred (1933/1986). 'The concept of truth in formalized languages'. In Tarski (1986), pp. 152–278

_____ (1944). 'The semantic conception of truth', *Philosophy and Phenomenological Research* 4: 341–75; reprinted in Blackburn and Simmons (1999), pp. 115–43.

_____ (1986). *Logic, Semantics, Metamathematics*, 2nd edn., ed. and introduced by John Corcoran, Hackett.

van Heijenoort, Jean (ed.) (1967). *From Frege to Gödel: A Source Book in Mathematical Logic*, Harvard University Press

This page intentionally left blank

INDEX