Revenge:

Dialetheism and its Expressive Limitations

Gareth Young

MA, Mlitt

Submitted in fulfilment of the requirements for

the degree of Ph.D.

Philosophy

School of Humanities

University of Glasgow

September 2014

Abstract

This thesis is about dialetheism and the problem of revenge. More broadly, it is about truth and what the logical paradoxes tell us about the logical behaviour of truth. One of the driving forces behind the contemporary study of truth and paradox is the problem of revenge: that many, perhaps all, available theories of truth, give rise to further paradoxes, invoking central notions of those theories, which demonstrate that the theory cannot express those notions. This sort of expressive limitation, especially if it involves the very notion invoked to diagnose what goes wrong in paradoxical sentences, would normally be thought a decisive point against a given theory of the paradoxes, were it not for the fact that the problem is so pervasive that every currently available theory has, at some point, been argued to suffer from it.

Dialetheism, the view that some contradictions are true, has often been thought to be the only view which has a reasonable chance of avoiding the problem. Indeed, the surge of interest in the view since the first publication of Priest's *In Contradiction*, in 1987, defending dialetheism, is probably due in large part to the seeming immunity to the revenge problem that Priest's view possesses. Its virtue, in respect of revenge, is that its ability to accept, without collapse into incoherence, contradictions, allows it to accept any further revenge paradoxes as merely giving more sound arguments for dialetheia (true contradictions).

This thesis argues that this appearance of revenge-immunity is mistaken. Dialetheism, too, has its revenge problems. The seeming virtue of dialetheism, that it can accept the contradictions generated by revenge paradoxes without incoherence, also has its drawbacks. This is because dialetheists are not only *able*, but *compelled* to accept the contradictions arising from the semantic paradoxes. This means that contradictions can arise in certain areas where they are undesirable. In particular, there are notions which seem to require *consistency* in order to be expressible. If we can demonstrate, using revenge paradoxes, that, on dialetheism, predicates putatively representing these notions would have to behave *inconsistently*, then we can demonstrate that dialetheists cannot express the notions.

There are many ways one might wish to carve up the different varieties of dialetheism available. I have separated the view into two broad kinds: *metatheoretically paraconsistent dialetheism*, on the one hand, and *metatheoretically consistent dialetheism*, on the other. This distinction decides to which variety of revenge problem the version of dialetheism in question is subject. I take each in turn, and argue that they are each subject to expressive limitations brought about by revenge paradox.

Revenge: Dialetheism and its Expressive Limitations

Contents

Acknowledgements

I would firstly like to sincerely thank my supervisors, Dr Adam Rieger and Professor Alan Weir. I am now approaching the end of the eleventh year of my time at the University of Glasgow, and have known Adam and Alan as friends and philosophical mentors for most of that time. I received a great deal of help and encouragement from both, to the great benefit of this thesis, and my philosophical development more generally.

All the members of the Department of Philosophy at Glasgow, staff and graduate students, have, at some point during my time here, and probably at some point during the last four years of my PhD, helped me in some way or other: either with helpful comments on my thesis, encouraging and critical questions in response to talks I have given, advice on my philosophical career, or general philosophical (or non-philosophical) discussion over pints of real ale. So I would like to express my gratitude to the staff at Glasgow, Dr David Bain, Professor Michael Brady, Dr Campbell Brown, Dr Ben Colburn, Dr Jennifer Corns, Dr Robert Cowan, Dr Victoria Harrison, Dr Gary Kemp, Dr Hugh Lazenby, Dr Stephan Leuenberger, Dr Chris Lindsay, Professor Fraser Macbride, Professor Fiona Macpherson and Dr Martin Smith. Also to the graduate students with whom I have studied, Carole Baillie, Umut Baysan, Stuart Crutchfield, John Donaldson, Suzanne Harvey, James Humphries, Dr Frederique Janssen-Lauret, Patrick Kaczmarek, Andrew MacGregor, Sheena McAnulla, Neil McDonnell, Stephanie Rennick, Catherine Robb and Abraham Sapien-Cordoba.

My family have been a source of unwavering support and encouragement, without which my philosophical studies would have been much more difficult, in numerous ways. I would like to thank, especially, my parents, John and Elizabeth. Special thanks are also due to my sister, Lisa Brown, whose fault it is, since it was her suggestion, that I changed my degree subject, from Computer Science and Mathematics, to Philosophy (which was, at the time, merely a fascinating side subject).

Thanks, also, are due to Graham Priest, whose brilliant philosophical work on dialetheism was the main inspiration for this thesis. In addition to inspiring much of my philosophical thinking over the last four years, his ingenious defence of

what, to many philosophers, is a quite outrageous view, has taught me that arguments, not incredulous stares, are required if we are to be justified in rejecting even the most counter-intuitive theories. He has also been very generous with his time, discussing various issues dialetheic in emails and in conversation, in a helpful and constructive way, despite my spending much of my thesis arguing against some of the central conclusions of his work. I would like to thank JC Beall, to whose work, *Spandrels of Truth*, I also devote a lot of discussion, which was helped by his generous comments by email and in conversation. I would also like to thank Franz Berto, who invited me to speak at the Northern Institute of Philosophy, and whose helpful discussion of my paper has clarified my thinking on issues of exclusion and revenge.

I would like to extend my deepest thanks to my partner, Julie McHugh, whose constant support and love has done much to make this thesis possible. Philosophers can be a difficult bunch, and she has tolerated me with great fortitude. Last, but definitely not least, I would like to thank Basil Carbonara McHugh, my miniature dachshund, who has provided constant amusement and companionship. If Basil could speak, I'm sure he would be a very wise philosopher, though we could not understand him.

University of Glasgow | College of Arts

Declaration of Originality Form – Research Degrees

This form **must** be completed and signed and submitted with your thesis.

Please complete the information below (using BLOCK CAPITALS).

Name .....
Student Number .....
Title of degree .....
Title of thesis .....

The University's degrees and other academic awards are given in recognition of a student's personal achievement. All work submitted for assessment is accepted on the understanding that it is the student's own effort. **Plagiarism** is defined as the submission or presentation of work, in any form, which is not one's own, without **acknowledgement of the sources.** For further information on what may be considered 'plagiarism', please read carefully the University's Statement on Plagiarism as contained in the University Calendar**.**

I confirm that this thesis is my own work and that I have:	
Read and understood the University of Glasgow Statement on Plagiarism	☐
Clearly referenced, in both the text and the bibliography or references, **all sources** used in the work	☐
Fully referenced (including page numbers) and used inverted commas for **all text quoted** from books, journals, web etc.	☐
Provided the sources for all tables, figures, data etc. that are not my own work	☐
Not made use of the work of any other student(s) past or present without acknowledgement. This includes any of my own work, that has been previously, or concurrently, submitted for assessment, either at this or any other educational institution.	☐
Not sought or used the services of any professional agencies to produce this work	☐
In addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations	☐

DECLARATION:

I am aware of and understand the University's policy on plagiarism and I certify that this thesis is my own work, except where indicated by referencing, and that I have followed the good academic practices noted above

Signed .....

Chapter 1: Introduction

This thesis argues that dialetheism, the view that dialetheia (true contradictions) exist, suffers from the problem of revenge. It discusses, primarily, the work of two philosophers: Graham Priest and JC Beall. Priest's work represents the most thoroughly-developed version of dialetheism currently available. It is also of a very thoroughgoing kind and, in particular, his dialetheism is described in dialetheic terms; which is to say, his dialetheism has a paraconsistent metatheory. Beall has developed a more moderate version of dialetheism which is described classically; that is, his dialetheism takes classical *ZF* as its metatheory. This difference in metatheory makes a great difference to the prospects each theory has for avoiding revenge. Beall's classical metatheory would appear to allow the construction of notions, in its metatheory, which, were they expressible in his object theory, would entail *triviality* (the truth of every sentence). This, at least *prima facie*, means that this view faces the revenge problem. Priest's view, with its paraconsistent metatheory, is different, since (assuming his theory is non-trivial), no such notions can be constructed. So Priest's metatheoretically paraconsistent dialetheism would seem to have better prospects for revenge-freedom than Beall's. In fact, as I argue, both theories face revenge problems.

Chapter 2 gives a characterisation of the problem in terms of two 'recipes' for formulating revenge: one for what I call 'formal revenge' and one for 'informal revenge'. What, exactly, the problem of revenge should be taken to be has been a matter of some controversy, so it is likely that the recipes I give will also be somewhat controversial, and may well need refinement if they are to offer a fully precise account of revenge. Nonetheless, my purpose in this chapter is only to give enough of an account of the problem, as I see it, to serve the main purpose of the thesis, which is to argue that dialetheism suffers from a revenge problem. The recipes should be enough for this. I also, in Chapter 2, discuss which sorts of problems should be understood as genuine revenge problems. In particular, I disagree with JC Beall that certain forms of revenge are 'too easy'.

I spend more time discussing metatheoretically paraconsistent dialetheism than the alternative. There are a few reasons for this. The first is that Priest has

written a lot more on the former than has been written on the latter, so there is more material to be engaged with. The second is that my revenge problem for Priest depends more on the precise details of his view than do my arguments against metatheoretically consistent dialetheism. Finally, though the views take different metatheories, they still share much in common, so devoting the same amount of space to both theories would have involved a lot of unnecessary repetition.

Chapters 3 and 4 discuss metatheoretically paraconsistent dialetheism. The first lays out the view and its motivations, drawing largely on Priest's articulation of it. The second argues that it suffers from revenge. I present a novel revenge problem concerning the notion of invalidity, arguing that metatheoretically paraconsistent dialetheism must treat validity inconsistently to such an extent that 'invalid' and 'just invalid' cease to be expressible.

Chapters 5 and 6 discuss metatheoretically consistent dialetheism. Again, the first lays out the view, paying special attention to the version defended by JC Beall as the clearest articulation of it. The second argues that it suffers from revenge. I don't present a new revenge problem here, instead arguing, contra Beall, that the view suffers from well-known revenge problems of both the formal and informal kind.

Chapter 2: What is Revenge?

2.0 Introduction

The problem of revenge is one of the central, driving forces behind the philosophical analysis of truth and its relationship with the paradoxes. This chapter discusses the problem of revenge and how it should best be understood. I follow the Beall's presentation of the issues in his *Prolegomenon to Future Revenge* (2007) fairly closely for a couple of reasons: firstly, he gives a very perspicuous treatment of the problems and my own characterisation of revenge will, in effect, be a refinement of his and, secondly, I wish to disagree with him on the important issue of which putative revenge problems should be taken to be genuine revenge problems. Beall argues that certain, standard formulations of revenge are, in fact, 'too easy revenge' and so may be dismissed as pseudo-problems. If Beall were correct, then a number of theories often thought to face revenge problems, in fact, would not, or at least the standard means by which it would normally be thought they face revenge would fail. One such theory is the metatheoretically consistent dialetheism defended by Beall in a number of places, but especially in his (2009). This theory is a particular focus of Chapter 4 of this thesis, and so whether common strategies for revenge are ruled out, for this theory, from the outset by a proper characterisation of revenge is of some importance.

First, I give a rough characterisation of how the problem of revenge arises, particularly in the context of classifying liar sentences. I then critically discuss Beall's 'too easy' objection to certain formulations of the revenge problem. Beall claims that the fact that the notion rendered inexpressible by revenge is classical means that we should expect it to trivialise in non-classical settings and, so, it is not a 'target' notion of our theory. Establishing the relevance of the inexpressibility result, he thinks, relies on the assumption, which may be rejected, that the semantics of our artificial language (which, we hope, models some relevant features of natural language) are intended to model the semantics of our natural language. Against Beall I argue that, firstly, it is not clear how one could make sense of the notion 'classically constructed' such that

it validates his claim that these notions should be expected to trivialise in non-classical languages; indeed, there are good reasons to think this is not the case. Secondly I argue that this is irrelevant anyway, since it doesn't matter whether a notion is a 'target' of the theory or not: what matters is that the model language avoids triviality differently from natural language i.e. via expressive limitations. Finally, I argue that, even if we accepted that the notion being 'target' is important, the way Beall suggests avoiding the relevance of the inexpressibility result doesn't help: he merely gives a promissory note for some further theory which we have good reason to suppose will be subject to a revenge problem that Beall has already granted is problematic rather than 'too easy'.

I then discuss Beall's characterisation of revenge via three recipes for revenge he provides. I suggest some modifications to these which seem to me to improve on them, most importantly in characterising revenge for the metatheoretically inconsistent dialetheism which is the focus of Chapters 3 and 4 of this thesis. There is some further critical discussion here of Beall's views on which recipes generate real revenge problems. I conclude the discussion by giving my own account of revenge in terms of two recipes: one for what I call 'formal revenge' and the other for what I call 'informal revenge'. Finally, I say a very few things about extending the treatment to forms of revenge arising from paradoxes other than the liar.

A caveat concerning the following account of revenge is this: my aim is to characterise, in a rough way, the problem of revenge as it is currently understood, and as best fits the rest of the discussion of revenge in this thesis. It is not my intention to give a historical account of the 'genealogy of revenge', though such a project would undoubtedly be interesting. What, exactly, revenge is supposed to be has been a controversial matter among philosophers. It is likely, then, that my own characterisation of the problem will prove somewhat controversial. It is also, as I have said, not utterly precise: I don't give a precise, formal account of what I take a liar paradox to be, for example, and I do not specify, exactly, what the relationship ought to be between the notions σ and λ (used in the recipes below), in the cases where it is not identity. Nonetheless, I hope, and think, that leaving the understanding of these notions at a more-or-less intuitive level will not detract from the purpose of this chapter, which is

simply to give enough of an understanding of the problem of revenge to follow the rest of the thesis.

A final caveat is that Beall, in his discussion, uses the term 'language' in a more substantial sense than is, perhaps, usual, seemingly intending it to refer to what might, more usually, be called a 'theory' (a language, in the purely syntactic sense, with the addition of a class of models interpreting it). I follow him in this usage throughout the thesis.

2.1 The Phenomenon of Revenge

A proper account of the paradoxes is central to our understanding of truth. One principle that truth clearly seems to obey and, in fact, which is plausibly constitutive of the very notion of truth is the T-scheme: $T\langle\alpha\rangle \leftrightarrow \alpha$. But combining this principle, along will some seemingly obvious truths about logic delivers inconsistency and, then, triviality. Consider the sentence L, equivalent to $\neg T\langle L\rangle$, and argue as follows:

(1) $T\langle L\rangle$	(suppose for *reductio*)
(2) L	(1, T-scheme)
(3) $\neg T\langle L\rangle$	(2, substituting equivalents)
(4) $\neg T\langle L\rangle$	(1, 3, *reductio*)
(5) L	(4, substituting equivalents)
(6) $T\langle L\rangle$	(5, T-scheme)
(7) $T\langle L\rangle \wedge \neg T\langle L\rangle$	(4, 6, $\wedge$-introduction)

Informally, we suppose that L is true and, since it says its untrue, conclude that it must be untrue. But since it's untrue, and says so, we conclude that it's true after all and, hence, both true and untrue. Since, classically, the principle of explosion, $A \wedge \neg A \vDash B$, is valid, the contradiction in (7) trivially entails everything[1]. Assuming we wish to reject the absurd result that everything is

[1] There are a number of versions of the *reductio* rule available, each of which, I think, would be equally suitable here. For example, Priest endorses a fairly weak *reductio* rule $\alpha \rightarrow \neg\alpha \vDash \neg\alpha$ in his (2006, p.87). It will do no harm to assume that, throughout the thesis, this is the *reductio* rule employed.

true, something in this argument must be rejected. It is normally thought (though not, of course, but dialetheists) that since the argument from the existence of *L* to (7) concludes with a contradiction that there is something wrong with this argument, so something must be given up and some diagnosis given for what it defective about *L*.

Since *L* cannot consistently be either true or false, many philosophers have found attractive the idea that *L* must then be neither true nor false. But, supposing the behaviour of 'neither true nor false' is as it seems, this raises a difficulty. If something is *neither* true *nor* false, then it seems obviously to follow that it is, *a fortiori*, not true. But if we claim that the liar is neither true nor false and are thereby committed to *L* being untrue, then no progress has been made, since we are back with contradiction.

The problem is general: we want to diagnose what is defective about the liar sentence *L* and so we place it in some category of defective sentences, say, to use Beall's terminology (2007, p.3), *bugger*. Whatever other principles we include in our account of buggerhood, we had better not have the principle *Bug*.

(*Bug*) $bugger\langle\alpha\rangle \rightarrow \neg T\langle\alpha\rangle$

If it follows from the fact that a sentence falls under the category *bugger* that the sentence is untrue, then we can construct a sentence (normally by techniques of diagonalisation), σ, equivalent to $bugger\langle\sigma\rangle \vee \neg T\langle\sigma\rangle$ and derive a contradiction in much the same way as with *L*. On the other hand, the whole point of introducing the notion *bugger* was to characterise liar sentences as defective in some way and so, presumably, as things we ought to reject. That all sentences of which *bugger* holds be untrue, therefore, seems essential.

The upshot of this is that any theory classifying liar sentences as *buggers* cannot have *bugger* as a notion expressible in the object language, since this would allow the construction of sentences like σ, which deliver inconsistency and then, by explosion, triviality. So the theory cannot express the notion *bugger*. But this is a disastrous expressive limitation, since the whole point of the theory is to classify liar sentences as *buggers*. This is the revenge problem for the theory which wishes to classify liar sentences as *buggers*: once the notion of *bugger* is introduced, the very liar reasoning the notion was introduced to diagnose

reappears and demonstrates that a crucial notion of the theory, viz. *bugger*, cannot be expressed in the theory. Revenge, in other words, is expressive limitation brought about by liar reasoning on the very machinery introduced to solve the paradoxes.

There are various objections that philosophers defending theories which classify liar sentences as defective might raise to the points made above, but my purpose is not to give the final word on the prospects for revenge-free versions of these theories, but to illustrate the shape of revenge problems and how they generally arise.

What we are interested in, when we construct theories of truth with an eye on the paradoxes, is building models (in the informal sense, though often formal model-theory is used) of natural language explaining how it is that the paradoxes do not deliver the triviality of natural language, despite seemingly sound arguments for that conclusion. So we construct a formal language, containing a truth predicate and various other salient features we think might hold of natural language (such as the inference principles we take to be valid and, perhaps, the T-scheme) and characterise the features of truth in this language. We hope, in so doing, to have given a picture of some of the features of natural language which explain how it is that liar reasoning does not collapse it into triviality.

To put the revenge problem in these terms, what revenge threatens to demonstrate is that we have given an inadequate model of natural language. Our model language cannot express (at least) one of its crucial notions, for example, *bugger*. If it could express this notion, triviality would follow; so the model language avoids triviality by making the notion inexpressible. Natural language, however, can express *bugger* (I have already expressed the notion in natural language several times). The model language is therefore a bad model of natural language. The model language achieves non-triviality by making a crucial notion like *bugger* inexpressible, but since natural language does not, the model sheds no light on how natural language achieves this feat.

The above sketch of the revenge problem has put it solely in terms of the revenge of the liar paradox. The restriction of discussion of revenge to the liar's revenge is almost universal in the literature, though this is probably due to the fact that the liar in general receives vastly more discussion in the literature than

the other paradoxes. There is no reason that similar revenge problems might not be developed from other paradoxes and, indeed, Stewart Shapiro (2007) has argued that the Burali-Forti paradox gives rise to revenge issues. As is standard, I focus my discussion of revenge on the liar's revenge, though I say a little towards the end of the chapter (section 2.4) about how one might easily accommodate other types of revenge arising from paradoxes other than the liar.

2.2 Too Easy Revenge

Following Beall's terminology, we let $\mathcal{L}_M$ be our artificial model language and $\mathcal{L}$ be our natural language (in the present case, English) which $\mathcal{L}_M$ is supposed to model. We abbreviate 'the behaviour of $\mathcal{L}_M$'s truth predicate' by '$\mathcal{L}_M$-truth' and the 'the behaviour of $\mathcal{L}$'s truth predicate' by '$\mathcal{L}$-truth'. So a revenge problem for $\mathcal{L}_M$ is the charge that $\mathcal{L}_M$-truth is a bad model of $\mathcal{L}$-truth, since $\mathcal{L}_M$ avoids triviality by lacking expressive powers present in $\mathcal{L}$. Since they avoid triviality in different ways, $\mathcal{L}_M$ does not tell us how $\mathcal{L}$ is non-trivial.

We can illustrate with a specific example given by Beall of a Kripkean partial-language approach (a nice account of which can be found in the appendix to Beall's paper (2007, pp.19-29)) where $\mathcal{L}_M$ is the fixed-point language of a Strong-Kleene construction. The metatheory of $\mathcal{L}_M$ is classical *ZF* set theory, in which we can construct notions like *true-in-*$\mathcal{L}_M$ and not-*true-in-*$\mathcal{L}_M$ to characterise $\mathcal{L}_M$-truth and for which it can be demonstrated that *a* is *true-in-*$\mathcal{L}_M$ if and only if *T*<*a*> is *true-in-*$\mathcal{L}_M$. We can then construct a sentence λ, equivalent to not-*true-in-*$\mathcal{L}_M$ <λ>. We can prove in the metalanguage that λ is *true-in-*$\mathcal{L}_M$ if and only if it is not-*true-in-*$\mathcal{L}_M$. Because the metatheory is classical, and so the law of excluded middle is a logical truth in it, we also have that λ is either *true-in-*$\mathcal{L}_M$ or not-*true-in-*$\mathcal{L}_M$, and hence that λ is both *true-in-*$\mathcal{L}_M$ and not-*true-in-*$\mathcal{L}_M$, which is a contradiction and, in the present context, delivers triviality. So, in order to prevent the construction of revenge paradoxes involving sentences like λ, *true-in-*$\mathcal{L}_M$ cannot be expressible in $\mathcal{L}_M$.

But the notion *true-in-*$\mathcal{L}_M$ is expressible in $\mathcal{L}_M$'s metatheory, as we have said, and this metatheory is contained in our natural language, $\mathcal{L}$. So *true-in-*$\mathcal{L}_M$ is expressible in $\mathcal{L}$. So, the revenge objection goes, $\mathcal{L}_M$-truth is a bad model of $\mathcal{L}$-

truth, since $\mathcal{L}_M$ avoids triviality by the inexpressibility of a notion expressible in $\mathcal{L}$. So, since they avoid triviality differently, $\mathcal{L}_M$'s non-triviality sheds no light on $\mathcal{L}$'s non-triviality.

Beall thinks, at least in this form, that the preceding, putative revenge problem is 'too easy revenge', and so not really a revenge problem at all (2007, p.10). What requires to be established to make it a genuine revenge problem is not merely the result that there are notions expressible in $\mathcal{L}$ which are inexpressible in $\mathcal{L}_M$, but also what is the *relevance* of this result.

A point against its relevance, according to Beall, is that *truth-in-*$\mathcal{L}_M$ is a classically constructed, model-dependent notion and so not one of the target notions we are aiming to model with $\mathcal{L}_M$. It is "hardly surprising", Beall thinks (2007, p.12), that such classically constructed notions trivialise is non-classical languages and so no surprise that the non-classical languages in question, such as $\mathcal{L}_M$, cannot express them.

One thing that would establish the relevance of the inexpressibility result, in Beall's view, is the assumption that the semantics of $\mathcal{L}_M$ are intended to model the semantics of $\mathcal{L}$. Since the semantics given for $\mathcal{L}_M$ involve, essentially, *truth-in-*$\mathcal{L}_M$, the semantics for our natural language, $\mathcal{L}$, must involve something similar. But since, as we have established, *truth-in-*$\mathcal{L}_M$ is not, on pain of triviality, expressible in $\mathcal{L}_M$, $\mathcal{L}_M$ is an inadequate model of the natural language $\mathcal{L}$, since $\mathcal{L}$ can express its own semantic notions, whereas $\mathcal{L}_M$ cannot.

Whilst this would, he thinks, turn 'too easy revenge' into a genuine revenge problem, it turns, crucially, on the assumption that the semantics of $\mathcal{L}_M$ are intended to be a model for the semantics of the natural language, $\mathcal{L}$. This assumption, according to Beall, may well be rejected. For example, suppose one thought that a proper account of the semantics of natural language is not to be given in terms of truth-conditions, but something else. Let's suppose, for definiteness, one thought semantics ought to be done in terms of use-conditions (see, for example, Field (2001) or Horwich (2005)). One would still, says Beall, face questions about the paradoxes and what, in light of them, we should take to be the logical behaviour of the truth predicate.

It would be quite reasonable, Beall thinks, to respond by giving truth conditional semantics, employing model theory, for a language $\mathcal{L}_M$, with the intention that the logical behaviour of the truth predicate in $\mathcal{L}_M$ model the logical behaviour of the truth predicate in $\mathcal{L}$, thus explaining how the paradoxes are to be avoided (or, perhaps, accommodated). In this case, according to Beall, the semantics for our language $\mathcal{L}_M$ are not intended to be a model for the semantics for $\mathcal{L}$, though $\mathcal{L}_M$-truth is intended to be a model for $\mathcal{L}$-truth. It would be no objection to this view, Beall thinks, that *truth-in-*$\mathcal{L}_M$ is inexpressible in $\mathcal{L}_M$ since that notion is employed in a merely instrumental fashion, in describing the semantics of the language $\mathcal{L}_M$, with the real goal being the modelling of $\mathcal{L}$-truth by $\mathcal{L}_M$-truth.

There are a number of problems here. Firstly, it is unclear what Beall might mean by 'classically constructed' which would support the claim that it is unsurprising that such notions trivialise in non-classical contexts. This objection is raised by Kevin Scharp in his (2013, pp.90-91). The obvious way of characterising a notion as classical is as some *n*-place predicate of which all the classical principles governing such predicates hold. To take Scharp's example, a classical notion of redness would be a one-place predicate such that everything is either red or not red, nothing is both red and not red etc. But there is nothing problematic, in general, with such notions existing in, and being defined for, non-classical languages. Indeed, as Scharp points out, since the theorems and inference principles of non-classical logics are generally proper subsets of those holding in classical logic, if one can't prove a contradiction from a notion classically, *a fortiori*, one can't prove a contradiction from the notion non-classically.

On the other hand, to use one of Scharp's examples again (2013, p.91), smooth infinitesimal analysis is a theory of infinitesimals consistent in intuitionistic logic, but inconsistent in classical logic. So the notion of infinitesimal implicitly defined in the theory is a non-classical notion which is inconsistent (and therefore trivial) in classical logic. So, though Beall says it is unsurprising that classical notions trivialise in non-classical contexts, this isn't right, at least on the present understanding of classical notion, things are the other way around: it is non-classical notions which trivialise on classical logic.

It may be that there is some other understanding available which would support Beall's claim, but it's not obvious what it would be. Here is one reason to think there can't be such a notion (or, at least, why Beall should hope there isn't). If one takes a non-classical model language, $\mathcal{L}_M$, to be an accurate model of our natural language, $\mathcal{L}$, then the correct logic for $\mathcal{L}$ is the non-classical logic which holds in $\mathcal{L}_M$; for example (see Chapter 5 for discussion), Beall endorses the logic *BXTT*, and so this ought to be the logic both of $\mathcal{L}_M$ and of $\mathcal{L}$. If there is an account of 'classical notion' such that it becomes obvious that these notions collapse into triviality whenever they are expressible in a non-classical context, then those notions had better not be expressible in $\mathcal{L}$, otherwise our natural language trivialises (because it has the same logic as $\mathcal{L}_M$). But, since $\mathcal{L}$ contains the (classical) metatheory of $\mathcal{L}_M$, any notion expressible in this classical metatheory is expressible in $\mathcal{L}$. Hence, if there are these classical notions which trivialise in non-classical settings like $\mathcal{L}_M$, it would seem to follow straightforwardly from their expressibility in $\mathcal{L}$ that natural language is trivial.

So Beall's claim that classical notions trivialise is non-classical settings, at the least, requires further development if it is to go towards supporting the claim that the inexpressible notion, *true-in-*$\mathcal{L}_M$, is not a target notion of the model.

A second problem is that it does not matter whether the notion is 'target' or not. We have a model language, $\mathcal{L}_M$, which has a certain logic, contains a truth-predicate, and so on. The defender of $\mathcal{L}_M$ as a model of $\mathcal{L}$ hopes that these are features $\mathcal{L}_M$ shares with our natural language, $\mathcal{L}$. We have a classical metatheory for $\mathcal{L}_M$, in which the notion *true-in-*$\mathcal{L}_M$ is expressible. We can demonstrate via liar reasoning and the properties of the metatheory that, if $\mathcal{L}_M$ can express *true-in-*$\mathcal{L}_M$, then triviality follows. So the expressibility of *true-in-*$\mathcal{L}_M$ in our object language, the non-classical logic of the object language and the classical principles of the metatheory are an explosive mix: they deliver triviality. This is avoided in the case of $\mathcal{L}_M$ by *true-in-*$\mathcal{L}_M$ being inexpressible. The model language, $\mathcal{L}_M$, is therefore non-trivial only because *true-in-*$\mathcal{L}_M$ is not expressible in it. How does our natural language, $\mathcal{L}$, achieve this? That is, how does it avoid succumbing to triviality through revenge paradox? We are none the wiser, because *true-in-*$\mathcal{L}_M$ is expressible in $\mathcal{L}$. Whether the notion whose inexpressibility allows the model language to be non-trivial is 'target' or not does not matter.

What is important is that it is the inexpressibility of this notion which prevents liar reasoning from trivialising $\mathcal{L}_M$; since $\mathcal{L}$ can express the notion, the model language, $\mathcal{L}_M$, has not shown us how $\mathcal{L}$ is not trivialised by liar reasoning. Hence, $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$.

A final problem for Beall's 'too easy revenge' objection concerns the instrumentalism he suggests as a way of avoiding the relevance of the inexpressibility result. He says that, were the semantics of $\mathcal{L}_M$ intended to model those of $\mathcal{L}$, the objection would be a genuine revenge problem. But, he thinks, we may well give up this assumption. In the example considered above, this might be done by someone who thought that semantics was a matter of giving, not truth-conditions, but use-conditions (for example, someone who wished to endorse the sort of use-theoretic semantics defended by Field (2001) or Horwich (2005)). One who thought this would still have to say something about the paradoxes and give an account of the logic of the truth predicate. What they might do, Beall suggests, is give, in an instrumental way, a truth-conditional semantics (using truth-in-a-model) for the object language $\mathcal{L}_M$, intending the account of $\mathcal{L}_M$-truth this provides to be a model for $\mathcal{L}$-truth, but not of the semantics of $\mathcal{L}$ more generally.

The first thing to point out about such a view is that it is very incomplete. The use-theoretic semantics which are intended to genuinely reflect the semantics of $\mathcal{L}$ haven't yet been given. We still need some model language, say $\mathcal{L}_U$, whose semantics are given use-theoretically and which, plausibly, provides a model of the semantics of $\mathcal{L}$. Moreover, these semantics, presumably, are not completely independent of truth and of logic. For one, it ought to be the case that $\mathcal{L}_U$ be developed such that the logic which holds in $\mathcal{L}_U$ and that holding of $\mathcal{L}_M$ be the same and, further, $\mathcal{L}_U$ ought to be able to contain a truth-predicate with the properties ascribed to the truth-predicate of $\mathcal{L}_M$ (it ought to obey the T-scheme, and so on). If this could not be done, the semantics modelled in $\mathcal{L}_U$ and the logical behaviour of the truth predicate modelled in $\mathcal{L}_M$ would be incompatible and so could not jointly hold of our target, natural language, $\mathcal{L}$, making at least one of them incorrect as an account of the latter. Supposing this could be done, I see no reason to suppose that our revenge objection could not simply be rerun in this new context. In particular, though this might depend on the details of the

theory, it seems extremely likely that there will be notions, such as *true-in-*$\mathcal{L}_U$, perhaps, which by reasoning parallel to the $\mathcal{L}_M$ case, we could demonstrate to be expressible in $\mathcal{L}_U$'s metatheory, but inexpressible, on pain of triviality, in $\mathcal{L}_U$. If it could be shown that there are no such notions for $\mathcal{L}_U$, it would be an extremely surprising and significant result, representing a great leap in the attempt to develop revenge-free accounts of truth and the paradoxes. On the other hand, it doesn't seem like the mere fact that the semantics of $\mathcal{L}_U$ is given use-theoretically is likely to deliver this result on its own. It may be that there are no notions constructible in the metatheory of $\mathcal{L}_U$ but inexpressible in $\mathcal{L}_U$, but given the resilience of liar reasoning, it seems sensible to withhold belief until we are actually presented with the theory.

If there are such notions, then the revenge problem generated would be genuine, and not too easy, since the semantics of the object language, in this case, *would* be intended to model the semantics of natural language. As Beall says of the case where this is granted of $\mathcal{L}_M$ (2007, p.10), we should conclude that $\mathcal{L}_U$ is an inadequate model of our real language, $\mathcal{L}$, since the latter can express its own semantic notions and the former cannot.

So, against one standard way of getting revenge on a theory of truth, Beall has accepted the inexpressibility result, but questioned its relevance, since the notion doesn't seem to be a target one. In response to Beall, I have argued that his invocation of the inexpressible notion's classicality does not help decrease the relevance of the result. I argued that his suggestion that, the semantics are treated instrumentally, as a tool for specifying the logic of the truth predicate, rather than a model for the semantics of natural language, faces the problem that, firstly, it renders the view unsatisfactorily incomplete and, secondly, that, assuming the view could be completed, exactly the same revenge problem would seem to recur. I also argued that, in any case, whether the notion rendered inexpressible in the object language is target or not is immaterial from the point of view of revenge. The revenge problem is that our model language achieves its features, especially non-triviality, via expressive limitations, whereas natural language does not. Whether these expressive limitations involve a 'target' notion does not matter.

2.3 Recipes for Revenge

In his (2007), Beall gives an account of the problem of revenge by giving three recipes by which revenge problems can be constructed. In this section, I discuss and refine Beall's recipes, finally offering two of my own, characterising the two main varieties of revenge problem: what I call 'formal revenge', on the one hand, and 'informal revenge', on the other. I also critically discuss, along the way, some remarks by Beall about which recipes are liable to generate genuine revenge problems.

Beall's recipes for revenge are as follows (2007, pp.11-12):

Rv1. Revenge Recipe 1.

We find some semantic notion, x, constructed in (and, hence, expressible in) the metalanguage of $\mathcal{L}_{M,}$ which is used to classify sentences of $\mathcal{L}_M$. We demonstrate that, on pain of triviality, x is not expressible in $\mathcal{L}_M$. We conclude that $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.

Rv2. Revenge Recipe 2.

We find some semantic notion, x, which, irrespective of whether it is explicitly used to classify sentences of $\mathcal{L}_M$, is constructed in (and, hence, expressible in) the metalanguage of $\mathcal{L}_M$. We demonstrate that, on pain of triviality, x is not expressible in $\mathcal{L}_M$. We conclude that $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features, especially non-triviality.

Rv3. Revenge Recipe 3.

We find some semantic notion, x, which is expressible in $\mathcal{L}$. We demonstrate that, on pain of triviality, x is not expressible in $\mathcal{L}_M$. We conclude that $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.

The only difference between Rv1 and Rv2 is in the constraints on what the semantic notion, *x*, is used for in $\mathcal{L}_M$'s metalanguage. In Rv1, the notion must be used in classifying the sentences of $\mathcal{L}_M$, whereas, in Rv2, this stipulation is absent. The point of this differentiation, presumably, though Beall does not say

so explicitly, is to ensure that revenge problems conforming to recipe Rv1 involve paradoxical constructions. Roughly speaking, liar paradoxes arise from notions classifying sentences by semantic values like 'true', 'false', 'untrue', and perhaps other notions like 'neither', 'meaningless', 'indeterminate' and so on. Given some classification of sentences under such notions, we construct sentences which try to classify themselves as falling under one of the categories other than truth, e.g. if the semantic values are true, false and neither, the relevant liar sentence would be α, equivalent to 'α is either false or neither'. The sort of revenge problem one might try to generate from such a liar paradox fits Rv1, since the notions 'false' and 'neither' are notions of semantic classification.

Since it is not part of Rv2 that the notion, ***x***, must be used for the classification of sentences of $\mathcal{L}_M$, it is not required that revenge problems fitting this second recipe involve paradoxical constructions like the liar. However, since what Beall says of ***x*** in his specification of Rv2 is "irrespective of whether it is explicitly used to classify $\mathcal{L}_M$-sentences [,it is expressible in $\mathcal{L}_M$'s metatheory]" (2007, p.11), neither is it precluded that ***x*** be used to classify $\mathcal{L}_M$-sentences and, thus, it is not precluded that paradoxical constructions count as fitting Rv2. The effect of this is that any putative revenge problem fitting recipe Rv1, *a fortiori*, is also an instance of Rv2.

When we try to get revenge on a theory of truth, we wish to find some notion expressible in our natural language, $\mathcal{L}$, but inexpressible in the object language, $\mathcal{L}_M$, of the theory and to argue, on this basis, that the object language is an inadequate model of natural language. The difference between Rv1 and Rv2, on the one hand, and Rv3, on the other, is that in the former two cases, our strategy for constructing the inexpressible notion goes via the metatheory of $\mathcal{L}_M$, whereas in Rv3, we are permitted to construct the notion directly in $\mathcal{L}$. Though we are permitted, on Beall's characterisation of Rv3, to construct ***x*** directly in $\mathcal{L}$, we need not, since all that is stipulated is that ***x*** be expressible in $\mathcal{L}$. But, since $\mathcal{L}$ contains both $\mathcal{L}_M$ and its metatheory, anything expressible in either of the latter is also expressible in $\mathcal{L}$. For this reason, any problem counting as an instance of recipe Rv1 or Rv2 will, *a fortiori*, count as an instance of Rv3. So, in fact, the set of Rv1 revenge problems is a subset of the set of Rv2 problems,

which is a subset of the set of Rv3 problems. Since part of what we want to do is distinguish different recipes for revenge, it will be useful to modify Rv1-Rv3 slightly to block this, so that each gives a distinct set of revenge problems.

A further point is that, although Beall aims to explicitly separate, via Rv1 and Rv2, revenge problems proceeding via the metatheory into those where the notion, ***x***, is used to classify sentences of $\mathcal{L}_M$ and those where ***x*** need not be used for this purpose, he does not draw this distinction in the other type of revenge problem he wishes to capture in Rv3. With recipes of this sort, we can establish the expressibility of ***x*** directly in $\mathcal{L}$, without first constructing the notion in the metatheory of $\mathcal{L}_M$. But Beall does not distinguish between the cases in which x's of this sort are used to classify sentences of $\mathcal{L}_M$ and the cases in which they are not used for this purpose. To demonstrate the parallels between the two strategies then, it will be useful to split Rv3 into two distinct recipes, along the lines of Rv1 and Rv2.

We can refine, slightly, Beall's recipes, with these two considerations in mind: that the recipes Rv1-Rv3 should not be such that (letting, for a moment, the name of each recipe be the name of the set of problems falling under that recipe) Rv1 ⊆ Rv2 ⊆ Rv3, and that a similar distinction to that drawn between Rv1 and Rv2 should be drawn between the parallel cases falling under Rv3. The revised recipes are as followed (with Rv1 omitted, since it is unchanged):

Rv2*. Revenge Recipe 2*.

*We find some semantic notion, **x**, constructed in (and, hence, expressible in) the metalanguage of $\mathcal{L}_M$, but which is not used to classify sentences of $\mathcal{L}_M$. We demonstrate that, on pain of triviality, **x** is not expressible in $\mathcal{L}_M$. We conclude that $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.*

Rv3*. Revenge Recipe 3*.

*We find some semantic notion, **x**, which is not constructed in the metatheory of $\mathcal{L}_M$, but which is expressible in $\mathcal{L}$ and is used to classify sentences of $\mathcal{L}_M$. We demonstrate that, on pain of triviality, **x** is not expressible in $\mathcal{L}_M$. We conclude*

that $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.

Rv4. Revenge Recipe 4.

*We find some semantics notion, **x**, which is not constructed in the metatheory of $\mathcal{L}_M$, but which is expressible in $\mathcal{L}$ and is not used to classify sentences of $\mathcal{L}_M$. We demonstrate that, on pain of triviality, **x** is not expressible in $\mathcal{L}_M$. We conclude that $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.*

To distinguish those recipes falling under Rv1 from those falling under Rv2*, I have made explicit the restriction that, whilst the notion, ***x***, in Rv1 is used to classify sentences of $\mathcal{L}_M$, ***x*** is precluded from doing this under Rv2*. To distinguish the first two, Rv1 and Rv2*, from the second two, Rv3* and Rv4, I have stipulated that the notion, ***x***, appearing in recipes of the kind Rv3* and Rv4 are *not* to be constructed in the metatheory of $\mathcal{L}_M$, though it should be expressible in $\mathcal{L}$. I have also separated Beall's Rv3 into two distinct recipes, Rv3* and Rv4, along the lines of the distinction drawn between Rv1 and Rv2, such that the semantic notion, x, occurring in those problems falling under Rv3* must be used to classify sentences of $\mathcal{L}_M$, whereas ***x*** must not be used for this purpose if the recipe is of kind specified in Rv4.

As I have said, though Beall is not explicit about it, we should take the conditions involving the usage of ***x*** to classify sentences as being aimed at capturing the fact that such ***x***'s are involved in the construction of liar paradoxes. That ***x*** be a notion used to classify sentences is specified only in recipes Rv1 and Rv3*, and so these are the only recipes aimed directly at capturing revenge problems involving liar-paradoxical reasoning. So arguments which are instances of Rv2 or Rv4 do not involve paradoxical reasoning at all: ***x*** is just some semantic notion, not involved in classifying sentences, which is expressible in $\mathcal{L}$ (in the case of Rv2, via its expressibility in the metalanguage of $\mathcal{L}_M$),but not in $\mathcal{L}_M$. But as I have sketched the problem above, and as it is understood in the literature, paradoxical reasoning, especially liar-paradoxical reasoning is *essential* to the problem of revenge: after all, the 'revenge' in question is supposed to be 'revenge of the liar paradox'.

For this reason, it seems to me, problems falling under recipes Rv2 and Rv4 are not *revenge* problems at all, but something else. This is not to say that the inexpressibility in $\mathcal{L}_M$ of some non-classifying semantic notion ***x***, despite its expressibility in $\mathcal{L}$, would be unproblematic. It may be that such an expressive limitation would be just as problematic for a theory as a revenge problem, though this might depend on what, exactly, the notion turned out to be. But the fact that the inexpressibility does not seem to have any strong connection to the liar, or to paradoxical reasoning in general, precludes it, in my view, from being counted as a revenge problem proper.

On a similar note, we might wonder whether the requirement that ***x*** be a notion which *classifies* sentences of $\mathcal{L}_M$ really gets a tight enough grip on the fact that revenge problems involve liar reasoning. It doesn't seem essential to the fact that a notion is used to classify sentences that liar paradoxes are constructible from the notion. It also doesn't seem like a classificatory notion, expressible in $\mathcal{L}$, which doesn't generate liar paradoxes, is thereby guaranteed to be expressible in $\mathcal{L}_M$. Perhaps there are some such notions which are not. This being so, there may be problems fitting Rv1 and Rv3* which don't directly involve any liar-paradoxical reasoning. If this were so, there would be problems counting as revenge under recipes Rv1 and Rv3*, but which would not count as revenge, properly understood. We might fix this by adding explicitly the requirement that the semantic notion, ***x***, be one from which liar sentences are constructible.

So, with Rv2 and Rv4 rejected as sources of genuine revenge problems, we can modify Rv1 and Rv3* to explicitly invoke liar reasoning as follows:

Rv1*. Revenge Recipe 1*.

We find some semantic notion, λ, constructed in (and, hence, expressible in) the metatheory of $\mathcal{L}_M$ and demonstrate that, were λ expressible in $\mathcal{L}_M$, we could construct a sentence, β, equivalent to $\lambda\langle\beta\rangle \vee \neg T\langle\beta\rangle$, from which we can derive the contradiction $\lambda\langle\beta\rangle \wedge \neg\lambda\langle\beta\rangle$, from which follows triviality. We conclude that λ is inexpressible in $\mathcal{L}_M$ and, therefore, that $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.

Rv3** Revenge Recipe 3**

We find some semantic notion, λ, which is not constructed in the metatheory of $\mathcal{L}_M$, but which is expressible in $\mathcal{L}$ and demonstrate that, were λ expressible in $\mathcal{L}_M$, we could construct a sentence, β, equivalent to λ<β> ∨ ¬T<β>, from which we can derive the contradiction λ<β> ∧ ¬λ<β>, from which follows triviality. We conclude that λ is not expressible in $\mathcal{L}_M$ and, therefore, that $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.

These formulations have the advantage that it is explicit that the inexpressibility of the relevant semantic notion is a consequence of liar reasoning. These recipes capture the form of a good many revenge problems, perhaps most of them. But they still do not have universal application. In particular, they do not seem to fare well capturing the revenge problems which have been alleged to afflict metatheoretically inconsistent dialetheism, which is the focus of Chapters 3 and 4 of this thesis. The problems are, in essence, as follows. Firstly, it is not always the notion directly used in the construction of the crucial liar sentence which ends up being inexpressible: sometimes the notion is one either constructed from, or crucially dependent on some feature (especially the consistency) of the notion invoked in the liar sentence. Secondly, the role of triviality in these revenge problems is less clear. I explain each of these connected points in turn.

For theories which aim at complete consistency (in both object and metatheory), revenge problems generally aim at demonstrating that some notion expressible in our natural language, $\mathcal{L}$, if expressible in our object language $\mathcal{L}_M$, leads to contradiction. Since the logic of both object and metatheory is likely to validate the principle of explosion, this contradiction would immediately trivialise the theory and so the notion giving rise to it must be inexpressible in $\mathcal{L}_M$.

For (dialetheist) theories which allow inconsistency in the object language, but have a classical metatheory, the situation is similar. Against such theories, revenge problems normally (though not always) aim at demonstrating that some notion expressible in our natural language, $\mathcal{L}$, if expressible in our object language, leads to contradiction not just in the object language, but in the metalanguage. Though the object language is, presumably, paraconsistent, the

metalanguage is not, and so, again, triviality follows, and so the notion must be inexpressible in $\mathcal{L}_M$.

For thoroughgoing dialetheist theories, which are inconsistent at the levels both of object language and metalanguage, things are slightly different. Because there is little prospect of demonstrating triviality from (almost) any contradiction (since they may be absorbed without triviality by both object and metatheory), a slightly different approach is used. A full development of revenge problems for such theories is carried out in Chapter 4 of this thesis, but the situation is, very roughly, this. We try to construct some notion, expressible in $\mathcal{L}$, which seems to have consistency built-in to the meaning of the notion itself, and demonstrate that the characterisation of the notion on this dialetheist view must lead to its inconsistency, if it is expressible in $\mathcal{L}_M$. Since, if the notion is to have the required meaning, it must be consistent, this demonstrates that the notion is inexpressible in $\mathcal{L}_M$. The notions invoked in such putative revenge problems are numerous, and include consistency, disagreement, non-dialetheia, just true and just false (See Chapter 4 for an account of the problems arising from non-dialetheia and just false, and Shapiro (2004) for the rest).

Sometimes, with these notions, the attempt at getting revenge is similar to the cases above, and fits one of the recipes Rv1* or Rv3**. For example, in the case of the notion 'just false', we construct a liar sentence *L*, equivalent to '*L* is just false'. From *L*, we can reason in the usual way and demonstrate that *L* is both just false and true (and so also not just false). Since, we are supposing, it is part of the meaning of 'just false' that it behave consistently, i.e. that sentences which are just false can't also be true, the notion is thereby inexpressible in our object language, $\mathcal{L}_M$. So the liar sentence, *L*, is constructed using the notion which we wish to demonstrate is inexpressible.

But this is not always the case. Consider the potential revenge problem for dialetheism that it cannot express disagreement (see Shapiro (2004) for a characterisation of this problem). This arises from the fact that, for dialetheists, sentences are inconsistent, but not incompatible, with their negations. For example, if someone says to a dialetheist 'α', and a dialetheist replies '$\neg\alpha$', they have said something inconsistent, but not incompatible, with α, for the

dialetheist who thinks $\neg\alpha$ is true may yet think α it true as well. In general, for any acceptable form of negation we might introduce into a dialetheist theory, say, *NOT*, it seems reasonable to think we can always construct a sentence φ, equivalent to *NOT*-*T*<φ>, from which we can deduce both *T*<φ> and *NOT*-*T*<φ>. Since, for any sort of negation a dialetheist might introduce, we can construct a liar sentence with it and demonstrate that it leads to inconsistency, dialetheists have no negation such that a sentence is genuinely incompatible with its negation. This being the case, so the objection goes, dialetheists cannot express disagreement, since disagreement essentially requires the ability to say something with which the thing you wish to disagree is incompatible.

The objection, then, is that we can express disagreement in our natural language, $\mathcal{L}$, but, since no negation operator is an incompatibility operator in $\mathcal{L}_M$, disagreement cannot be expressed in $\mathcal{L}_M$, and so $\mathcal{L}_M$ is not a good model of $\mathcal{L}$. The important difference, here, between the 'just false' problem and the disagreement one just sketched is that, with the 'just false' problem, the notion we use to construct the liar sentence is the one we wish to demonstrate inexpressible. With the disagreement problem, the notion of disagreement does not occur in the liar sentence itself: instead we construct a standard liar sentence, for a given negation, and this demonstrates that sentences are compatible with their negation. The reason this (if the argument is correct) makes disagreement inexpressible is that disagreement essentially depends on the behaviour, in particular, the consistency, of negation.

So, in the case of disagreement, it is not the notion involved *directly* in the construction of the liar sentence which is shown to be inexpressible. Rather it is some notion which essentially depends on the behaviour of this notion for its expressibility[2]. Since the recipes above require that the inexpressible notion be such as to allow a liar sentence to be constructive *using that very notion*, the recipes don't capture this particular revenge problem. For another example, I try to show, in this thesis, that dialetheists cannot express the notion

[2] One might respond that it is an essential feature of negation that it expresses incompatibility and so the above argument *does* establish the inexpressibility of one of the notions in the liar sentence: namely, negation. In the present context, however, this would be question-begging. The central thesis of dialetheism is that some sentences can be such as to be true and have a true negation. So insisting that negation must always express incompatibility is just to insist that dialetheism is false. This doesn't mean it's question-begging to insist on some way of expressing incompatibility sometimes.

‘invalidity’, and that this is a revenge problem. Chapter 4 has the details on exactly how this is supposed to be achieved, but the liar sentence I construct simply says of itself that it is untrue in a certain model, *M*, and so does not contain the notion of invalidity which it is supposed to demonstrate inexpressible.

Similar points apply to other potential revenge problems for dialetheism, for example those involving the notions ‘non-dialetheia’ and ‘consistency’. Nonetheless, since the problems involve allegedly problematic expressive limitations arising from liar reasoning, they are clearly still revenge problems, and so our account should capture them.

To pre-empt an objection, one might imagine a dialetheist responding to this situation by insisting that, unless it is the very notion from which the relevant liar sentence is constructed which is inexpressible, the problem doesn’t count as genuine revenge. My response to this is twofold: firstly, it is very unclear what the motivation would be for such a stipulation, beyond the unsatisfactorily *ad hoc* attempt to avoid ones expressive limitations being classified as revenge problems. The central, basic feature of revenge is that attempted solutions to the liar paradox fall victim to expressive limitations forced on the theory by liar reasoning, which is exactly what happens in the case, for example, of the problem of disagreement. Secondly, this is a merely terminological quibble: nothing turns on whether we use the word ‘revenge’ for the problems in question. If a dialetheist was insistent on using the restriction under consideration, they could read instances of ‘revenge’ as ‘revenge-like’, if they wished, and the philosophically important points I make in the thesis would be essentially unchanged.

So, a revision to the revenge recipes such that the notion demonstrated to be inexpressible by liar reasoning is not required to be the very notion employed in the construction of the liar sentence itself, is required. The second difficulty which would seem to prevent Rv1* and Rv3** capturing the revenge problems supposed to afflict metatheoretically inconsistent dialetheism is the stipulation in those recipes that the semantic notion, *x*, be inexpressible *on pain of triviality*.

This point is related to the previous. Views on which at least some part of the theory, either the metatheory or both the object and metatheory, is consistent (with a logic validating explosion) lead to triviality straightforwardly, on the assumption that certain notions are expressible in the object theory. Not so for dialetheism with an inconsistent metatheory. On this view, since neither the object theory nor the metatheory supports explosion, almost no contradictions lead to triviality (the exceptions being ones involving sentences like 'Everything is true'), wherever they are expressible. So, if such theories suffer revenge problems, as I argue they do, they are not captured by recipes Rv1* and Rv3**, since those recipes stipulate that the contradiction delivered by the expressibility of the notion, ***x***, lead to triviality.

Consider, as an example, the notion 'just false', which metatheoretically inconsistent dialetheists are alleged to be unable to express. The notion, it is argued, must behave consistently, if it is to mean what is required (since, otherwise, some sentences which are just false are also true). But, for a dialetheist, a predicate aimed at representing the notion must behave inconsistently; because of the liar sentence which says of itself that it is 'just false'. Grant that this problem is genuine, and the notion really is inexpressible. It does not seem to be the case that, were the notion expressible, triviality would follow. The notion is simply inexpressible, since, once we introduce a predicate meeting an intuitive definition of 'just false' (presumably, 'false and not true'), this notion can immediately be shown to be inconsistent, and so not to mean 'just false', even though triviality does not seem to follow from it.

Shapiro's way of putting the objection is that the dialetheist cannot express the notion "unless the meta-theory is (completely) consistent" (2004, p.338). But this doesn't seem quite right. Replacing the dialetheists inconsistent metatheory with a consistent (presumably classical) one would not help the inconsistency of the notion defined by 'false and not true' in the object language, since the relevant liar argument would still go through in that context. Certainly, we could construct a consistent notion of 'just false' in the classical metatheory, but this would, were it expressible in the object theory, deliver triviality and so would simply be another revenge problem and 'just false' would still be inexpressible.

It seems to me that the correct way to revise the recipes is simply to drop the requirement that the notion be inexpressible in the object theory *because otherwise triviality would follow*, and simply require that the notion be inexpressible. After all, it is the inexpressibility of the notion, not the fact that it would entail triviality, which is important. The fact that a notion would entail triviality were it expressible in the object language is simply one way of establishing that the notion is inexpressible, but it is not, or need not, be the only way. If one can give good arguments that a notion has consistently as part of its meaning, then the fact that one can demonstrate its inconsistency on dialetheist theories is enough to establish its inexpressibility, whether it goes on to trivialise or not.

One might be concerned, here, that by removing the stipulation that the notion would entail triviality, were it expressible, undermines what was said in this chapter about revenge problems demonstrating that an object language, $\mathcal{L}_M$, is a bad model of a natural language, $\mathcal{L}$, because they *avoid triviality differently* – the former by expressive limitation, the latter not. But a little thought should assuage this worry.

With the original liar paradox, we have a simple argument, outlined in Section 2.1, from a liar sentence to triviality. We presume that our natural language is not trivial, and so, somehow, this argument to triviality fails in natural language. The job of our theories of truth is to provide model languages in which this argument fails which, we hope, show how natural language achieves this. If it is an essential part of our model language that it be unable to express notions which are expressible in our natural language, then we do not have a good model of how natural language avoids the argument to triviality from the liar. The fact that, on metatheoretically inconsistent versions of dialetheism, triviality does not follow when we add a notion defined by 'false and not true' to the language does not matter. The important thing is that, if the revenge objection is right, and consistency is part of the meaning of 'just false', then the model language cannot express the notion, and so is crucially different to natural language. It's an essential feature of the account of how the model language avoids the argument to triviality from the liar that liar reasoning demonstrates it to be unable to express crucial semantic notions which are

expressible in natural language. The model language, therefore, sheds no light on how the trivial conclusion of the argument is avoided in natural language.

One final issue to mention, before giving the final versions of each recipe for revenge is which of the notions involved must be semantic. In recipes Rv1* and Rv3**, the revenge problem is supposed to arise from a single notion, λ, which is stipulated to be semantic (since it is meant to be involved in the construction of a semantic paradox). We are now allowing that there may be two distinct notions involved in revenge: one from which we construct a liar sentence and one, dependent upon the first, which is rendered inexpressible by liar reasoning. The first, it seems clear, ought to be a semantic notion, as in the previous recipes. It doesn't seem to me that we ought to require this of the second notion, though, of course, it should be permitted to be semantic. For an example which suggests this, consider again the disagreement problem. Arguably, disagreement is not a semantic notion; it may, instead, be an *attitude*, or an *action*, which we wish to use the language to express. Someone who thought this may still, quite reasonably, think of the fact (if, indeed, it is a fact) that dialetheists cannot express disagreement as a revenge problem. This being so, we should require that the notion employed in the construction of the crucial liar sentence be semantic, but relax this restriction for the second notion and allow that it need not be semantic. These considerations, finally, give the following two recipes for revenge:

RvF. Recipe for Formal Revenge.

We find some semantic notion, λ, constructed in (and, hence, expressible in) the metatheory of $\mathcal{L}_M$ and demonstrate that, were λ expressible in $\mathcal{L}_M$, we could construct a sentence, β, equivalent to $\lambda\langle\beta\rangle \vee \neg T\langle\beta\rangle$, from which we can derive the contradiction $\lambda\langle\beta\rangle \wedge \neg\lambda\langle\beta\rangle$. We demonstrate that this establishes the inexpressibility in $\mathcal{L}_M$ of some notion, σ, which is expressible in $\mathcal{L}$, where it is permitted (indeed, it is common) that $\lambda = \sigma$. We conclude that $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.

RvI. Recipe for Informal Revenge.

We find some semantic notion, λ, which is not constructed in the metatheory of $\mathcal{L}_M$, but which is expressible in $\mathcal{L}$ and demonstrate that, were λ expressible in $\mathcal{L}_M$, we could construct a sentence, β, equivalent to λ<β> ∨ ¬T<β>, from which we can derive the contradiction λ<β> ∧ ¬λ<β>. We demonstrate that this establishes the inexpressibility in $\mathcal{L}_M$ of some notion, σ, which is expressible in $\mathcal{L}$, where it is permitted (indeed, it is common) that λ = σ. We conclude $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.

These recipes compare favourably with Beall's. Firstly, we have jettisoned the recipes not involving paradoxical constructions and focused on the two that do as providing real revenge problems. Secondly, the importance of the liar paradox as the source of inexpressibility has been made explicit. The stipulated in RvI that *λ* not be constructed in the metatheory of $\mathcal{L}_M$ also means that instances of RvF are not trivially instances of RvI. Finally, allowing for the separation of the inexpressible notion from the one directly involved in the construction of the liar sentence, as well as the relaxation of the requirement that expressibility of the relevant notions must lead to triviality, has allowed the recipes to capture putative revenge problems for (metatheoretically inconsistent) dialetheism, which might otherwise have been missed.

I now discuss some of Beall's general remarks (2007, pp.12-14) about which recipes may be expected to generate real revenge problems. He says of his own Rv1 and Rv2 (corresponding roughly to my own RvF) that how one ought to respond to the charge of revenge will depend on the notion ***x*** (on his own recipes, but *λ* and *σ*, on mine), which seems fair enough. In general, however, he thinks the charge of inadequacy difficult to substantiate in these cases, even when we have established the inexpressibility result.

The reason he gives is that classical logic, normally, will be an extension of the logic of the model language, $\mathcal{L}_M$ (assuming, of course, that the logic of $\mathcal{L}_M$, as in the examples above, is not itself classical). This being the case, he thinks, it is perfectly acceptable that $\mathcal{L}_M$ be constructed in a classical metalanguage. In this context, it is entirely unsurprising that classically-constructed notions trivialise in the non-classical context of $\mathcal{L}_M$. According to Beall, this point is "often sufficient to blunt, if not undermine, a revenger's charge" (2007, p.12), if the

problem is constructed as in Rv1 or Rv2 (or RvF). What must be done to construct a genuine revenge problem is to establish, not just the unsurprising result that there are notions inexpressible in $\mathcal{L}_M$, but also the relevance of this result.

This is the objection already discussed, and my response is the same. Firstly, that a notion is classical does not make it unsurprising that it trivialises in non-classical languages; things are the other way around. Secondly, it does not matter whether a notion is a target of the theory or not. What matters is that the model language avoids the paradoxes by being expressively limited in ways natural language is not. The extent to which such a result is to be expected is not relevant either, since it is not part of any account of revenge that the expressive limitation established be surprising. A serious problem is no less serious for being obvious.

Beall then discusses attempts at revenge conforming to his Rv3, corresponding to my own RvI. Here, Beall says that he is assuming that the semantic notion, *x*, is used "neither explicitly not implicitly…for purposes of classifying sentences". This is not, strictly speaking, in keeping with the exact letter of Rv3, which does not stipulate that *x* not be used in this way. So it appears that Beall understands Rv3 along the lines of my own Rv4, rather than in the way it is actually phrased.

The task of demonstrating that a problem fitting Rv4 is really a revenge problem is, he says, "formidable" (2007, p.13), since it must be shown that *x* is really an intelligible notion of $\mathcal{L}$. Another reason the task is formidable, as should be clear now from my own discussion, is that problems fitting Rv4 are not really revenge problems, since they don't involve the liar paradox.

He refers the reader to his discussion of 'incoherent operators' (2007, pp.5-6) where he argues that there cannot be languages containing operators meeting all of a set of conditions. The conditions he gives are:

F1. The language contains a truth predicate, $T\langle x\rangle$, which obeys, unrestrictedly, the T-scheme (at least in its rule form)

F2. Reasoning by cases is valid

E1. $\vDash a \vee \varphi a$

E2. $a, \varphi a \vDash \bot$

If F1 and F2 hold, then the language cannot contain an operator, φ, satisfying E1 and E2 (at least, it can't contain an operator if the language is non-trivial). This isn't quite right, since the arguments from these conditions to triviality require certain standard structural rules to hold in the logic in question. Some philosophers and logicians have thought defended the idea that at least some of these structural rules must be rejected; for example, Weir (2013) has suggested that the transitivity of entailment be given up, Zardini (2014) that we give up structural contraction and Ripley (2013) that we give up structural cut. So, if we discover, of some putative operator, satisfying each of the above principles, it does not immediately follow that there is, in fact, no such operator in the language: it may be that the language is trivial or, less radically, that some structural rule of the logic, despite appearances, fails.

Beall calls an operator, φ, satisfying each of E1 and E2, an EE device. What the paradoxes teach us, according to defenders of paraconsistent or paracomplete theories, is that language cannot contain an EE device. If a revenge objection presents such theorists with a semantic notion, which is an EE device, and argues that, if expressible in $\mathcal{L}_M$, triviality follows, this may, Beall thinks, be question-begging. The views in question think that the lesson of the liar paradox is that languages cannot contain EE devices, and so a semantic notion, x, which is such a device, must, on these views, be incoherent. They may then simply insist that the claim that English, for example, contains such a device is question-begging. It may be, he thinks, that some argument could be given that the expressibility of the notion is important for some theoretical purpose. If so, this might settle things. But he is pessimistic about the possibilities of finding such arguments.

Beall is quite right that there are special difficulties with the sort of revenge problem which does not construct the semantic notion directly in $\mathcal{L}_M$'s metatheory. If we try to get revenge on a theory by constructing a notion in its metatheory, as in Beall's Rv1 or my RvF, then the properties of that notion are not in question, since the notion has been precisely and formally defined. If, however, the notion is one taken directly from our natural language, $\mathcal{L}$, then its exact properties may be a matter of reasonable dispute. For example Shapiro,

who charges that dialetheists cannot express the notion ‘just false’ (2004), insists that the notion must behave consistently. Priest, who responds on behalf of dialetheism, thinks otherwise. This makes establishing the inexpressibility of ‘just false’, more difficult, since whether it is expressible will likely depend, in part, on whether consistency is essentially part of the meaning of the notion.

But perhaps one source of Beall’s pessimism about finding notions which one can reasonably claim a theory ought to be able to express, and so may not simply dismiss as incoherent, is his reading of Rv3 as Rv4 i.e. on his assumption that the semantic notion in question not be one which classifies sentences of $\mathcal{L}_M$. The vast majority of any theorist’s pronouncements about their object language, $\mathcal{L}_M$, are not made in the formal metalanguage of $\mathcal{L}_M$, but in natural language. There are, therefore, in any philosopher’s characterisation of truth, a multitude of natural language sentences classifying sentences of $\mathcal{L}_M$ using notions from natural language. It is to these claims that we ought to look to find revenge problems fitting RvI. As Shapiro says in his attack on Priest’s metatheoretically inconsistent dialetheism (2004), the prospects of constructing notions in his formal theory, and showing that they deliver triviality, are not good, so “[p]erhaps the opponent can attack the dialetheist's *informal* remarks… to show…that the dialetheist is subject to a criticism much like one that Priest levels against consistent theories of truth: there are certain notions and concepts that the dialetheist invokes (informally), but which she cannot adequately express.”

This is the reason recipe RvI is a recipe for ‘Informal Revenge’, as opposed to the ‘Formal Revenge’ of RvF. The best way to get revenge of this kind it to look to a given theorists *informal* remarks in $\mathcal{L}$ about their object language, $\mathcal{L}_M$, with the goal of finding notions in $\mathcal{L}$ which cannot be expressed in $\mathcal{L}_M$. Since the notion in question has been explicitly invoked by the theorist, this makes it much more difficult for them to simply reject the notion as incoherent. The more pervasive the notion’s use, the more difficult this becomes. This is exactly Shapiro’s strategy in his (2004): all dialetheists regularly discuss consistent theories, as opposed to inconsistent ones, as well as the ‘normal’ sentences, like those which are false only, true only or non-dialetheia. Shapiro argues that the dialetheist cannot express these notions. But since the dialetheist is committed,

via their informal remarks, to the legitimacy of these notions, they cannot respond by simply dismissing those notions as unintelligible. The fact that we can appeal to notions employed by theorists informally gives reason to be much more optimistic about the prospects for generating revenge problems, as in RvI, by direct appeal to a notion's expressibility in $\mathcal{L}$ (rather than via $\mathcal{L}_M$'s metatheory) than Beall allows.

2.4 Varieties of Revenge

One final issue I wish to mention is the centrality of the liar paradox to the account of revenge given in this chapter. Though the liar paradox receives the lion's share of the discussion in the paradox literature, and is almost the sole focus of the literature on revenge, it may well be that revenge-like problems arise from other paradoxes. Stewart Shapiro, for example, has argued that the Burali-Forti paradox "has its own annoying revenge issues" (2007, p.321). If we wish to accept these as genuine revenge problems, it seems to me that the best way to respond to this different variety of revenge is to classify revenge problems by the paradox giving rise to them. So the above recipes, RvF and RvI characterise the revenge of the liar paradox, and two separate conditions describing formal and informal recipes for Burali-Forti revenge could also be given. The changes being that we drop the requirement that the notion λ be semantic (because the paradox is not a semantic one) and stipulate that the reasoning leading to the contradiction be of the same kind which appears in the Burali-Forti paradox, however that is best described. I see no reason that a similar treatment could not be extended to the revenge of other paradoxes, should such problems arise.

2.5 Chapter Conclusion

In this chapter, I have critically discussed some remarks by Beall on revenge. I have settled on two 'recipes' for revenge as describing the revenge problem's essential characteristics, given by RvF and RvI. These may well not be definitive, but they should give us enough of a grip on the problem for the purposes of the revenge arguments given in this thesis. I have characterised revenge as coming in, essentially, two forms: Formal Revenge and Informal Revenge. With the

former, formal sort of revenge problem, a notion is constructed in the formal metatheory of an object language $\mathcal{L}_M$, and it is argued that this notion, because of liar-paradoxes, is inexpressible in $\mathcal{L}_M$ and that, thereby, $\mathcal{L}_M$ is a bad model of natural language. With the latter, informal, sort of revenge, a notion is found in natural language (generally in the informal remarks of the theorist on whose view we wish to get revenge) and argued, again via liar reasoning, to be inexpressible in $\mathcal{L}_M$ and that, thereby, $\mathcal{L}_M$ is a bad model of natural language.

JC Beall gave some reasons to think both such strategies problematic. In the formal case, the inexpressible notion constructed in the metatheory, due to its being 'classically constructed' may, pending argument, be thought 'too easy' and, therefore, irrelevant. I have argued that this is not so; the problems, irrespective of the ease with which inexpressibility is established, are genuine revenge problems. In the informal case, Beall has said that it is likely that such natural language notions will simply be dismissed as incoherent, and the insistence that they give rise to revenge, question-begging. In response to this claim, I have suggested that the place to look for revenge notions in natural language is the informal claims of the theorist on whose view we want revenge. If we can construct revenge problems from notions appearing in these claims then, especially if the notion occurs pervasively, it is very difficult for the theorist to dismiss the notion as incoherent. This strategy looks more hopeful than Beall suggests.

Chapter 3: Metatheoretically Paraconsistent Dialetheism

3.0 Introduction

This chapter addresses the dialetheism of Graham Priest. His view is notable for a number of reasons. Priest's *In Contradiction* (2006), first published in 1987, is the first rigorously-logical book-length treatment of dialetheism. His first publication on the topic was his (1979) paper *Logic of Paradox*. So Priest's work is of historical importance as the first to take the view seriously as a solution to a number of puzzles, the most important of which being the semantic and set-theoretic paradoxes.

Priest has published extensively on the topic in the intervening years, refining his view in a second edition of *In Contradiction* (2006), four further monographs (*Beyond the Limits of Thought* (2002), *Towards Non-Being: the Semantics and Metaphysics of Intentionality* (2005), *Doubt Truth to be a Liar* (2006) and, most recently, *One* (2014)) and dozens of published articles. This makes Priest's dialetheism the best developed version of the view currently available.

A further notable feature of Priest's view is particularly relevant for our purposes: his dialetheism is extremely thoroughgoing. Another prominent dialetheist, JC Beall, defends the view that dialetheia occur only in semantic contexts (essentially, the only dialetheia there are contain the truth predicate). They do not, for Beall, occur 'in the world', in mathematics and, crucially, they do not occur in Beall's metatheory. Priest, on the other hand, defends dialetheist accounts of change, vagueness, certain puzzles in jurisprudence, the set-theoretic paradoxes and several other problems. Most importantly, from the point of view concerned with revenge, Priest's metatheory is paraconsistent. This allows Priest to treat his metatheory as contained in the object theory and gives it, perhaps, the best claim to revenge immunity of any current version of dialetheism.

A theory is subject to the problem of revenge if there is some notion expressible in natural language which we can use liar reasoning to show is not expressible in the theory's object language. Perhaps the most common strategy for

demonstrating that a theory has a revenge problem is to construct some notion in the theory's metalanguage which would generate triviality were it expressible in the object language. Since the metalanguage is part of natural language, the notion is expressible in natural language and the theory has a revenge problem. Since, on Priest's view, the metalanguage is contained in the object language, there can be no notion expressible in the former which is not expressible in the latter. So, on the assumption that Priest's theory is not trivial, there is no notion expressible in the metalanguage which would trivialise, were it expressible in the object language. This makes life more difficult for someone who wishes to argue that his view suffers from revenge, since it blocks the main strategy normally used.

3.1 Priest's Metatheoretically Paraconsistent Dialetheism

Dialetheism is the view that there are dialetheia: that is, true contradictions. A true contradiction is here understood, following Priest (2006, p.4) to be 'any true statement of the form: α and it is not the case that α'. Since it's generally accepted that falsity is truth of negation, one might equivalently say that a true contradiction is a statement which is both true and false.

This section presents Priest's view and the motivations he gives for it, primarily in the second edition of *In Contradiction* (2006). I follow Priest's presentation of his ideas closely, beginning with his discussion of the semantic paradoxes, followed by the set-theoretic paradoxes, before moving on to the more constructive discussion of the details of dialetheist logical theory including his views on truth, falsity, the connectives and entailment. I then move on to discuss Priest's account of paraconsistent set theory (which he takes as his working metatheory) and the recapture of classical reasoning.

Though dialetheists, such as Graham Priest, defend the existence of dialetheia in a number of areas, it is fair to say that the most powerful, and widely discussed, defence of the view concerns the semantic and, to a lesser extent, set-theoretic paradoxes. The central argument, at Priest sees it, is that since no consistent theory has succeeded, or could succeed, in resolving these paradoxes, an inconsistent (dialetheist) solution is the only option.

The separation of the logical paradoxes into the two categories ‘semantic’ and ‘set-theoretic’ is, for Priest, merely a matter of convenience (2006, p.9). Though the distinction is common, he thinks it almost impossible to draw the distinction satisfactorily. One reason for this is the existence of a formal isomorphism between the abstraction scheme of set theory and the Tarski satisfaction scheme:

$x \in \{y \mid \alpha\} \leftrightarrow \alpha(y / x)$

x satisfies $\langle\alpha\rangle \leftrightarrow \alpha(y / x)$

Where α is a formula containing only y free and $\alpha(y / x)$ is the result of substituting in α all free occurrences of ‘y’ with ‘x’, where and ‘$\langle\alpha\rangle$’ is a quotation name for α. Given this isomorphism, a number of the paradoxes in one category appear to have counterparts in the other. For example, Grelling’s paradox (which arises when we ask whether ‘heterological’ – where an adjective is heterological just if it does not apply to itself – is heterological) and Russell’s paradox (which arises when as ask whether the set containing all non-self-membered sets is self-membered) can each be transformed into the other, under this isomorphism.

However, according to Priest, not all paradoxes have such natural counterparts. For example, the Burali-Forti paradox does not have a natural semantic counterpart and the paradoxes of definability have no obvious set-theoretic counterpart. This is one reason he wishes to separate the two categories. A second is that the set-theoretic paradoxes have a solution which is (fairly) widely agreed-upon, whereas the semantic paradoxes do not.

In fact, in later work, Priest has given a precise account of the way in which the paradoxes are related: the familiar paradoxes such as the Liar and those from set theory are *inclosure paradoxes*.

3.2 The Semantic Paradoxes

Priest’s strategy is to give a set of conditions (derived from those given by Tarski (1936)) which, if satisfied by a language, are sufficient for its inconsistency. He

then defends the claim that natural language satisfies these conditions. The purpose of this strategy is not only to provide evidence that the paradoxes cannot be consistently solved, but also to explain why this is so.

According to Tarski, the paradoxes result from the supposition that language satisfies a set of closure conditions. Restricting, for simplicity, to the case of formulae containing one free variable, the conditions are as follows:

(1) For every formula, α, there is a term of the language, $\langle\alpha\rangle$, which names it.

(2) There is a formula with two free variables, *Sat*(*x y*), such that every instance of the scheme

$$Sat(\mathrm{t} \langle\alpha\rangle) \leftrightarrow \alpha(v/t)$$

is a theorem, where *t* is a term, α is any formula with one free variable, *v*, and $\alpha(v/t)$ is α with all free occurrences of '*v*' replaced by '*t*' (where we add the stipulation that '*t*' cannot occur free within the scope of a quantifier binding *v*).

(3) The rule of inference $\{\alpha \leftrightarrow \neg\alpha\} \vdash \alpha \wedge \neg\alpha$ is valid in the logic underlying the theory.

To show that these conditions are jointly inconsistent, we substitute for α the formula $\neg Sat(v\ v)$, and, for *t*, $\langle\neg Sat(v\ v)\rangle$. The scheme in (2) then delivers

$$Sat(\langle\neg Sat(v\ v)\rangle, \langle\neg Sat(v\ v)\rangle) \leftrightarrow \neg Sat(\langle\neg Sat(v\ v)\rangle\ \langle\neg Sat(v\ v)\rangle)$$

By the inference principle in (3), we then have

$$Sat(\langle\neg Sat(v\ v)\rangle, \langle Sat(v\ v)\rangle) \wedge \neg Sat(\langle\neg Sat(v\ v)\rangle\ \langle\neg Sat(v\ v)\rangle)$$

which is a contradiction. According to Priest, these conditions apply not only to certain formal languages, but to natural language, though not in exactly this form. Substituting the purely formal notions in the above conditions for natural language counterparts suggested by Priest, we get the following conditions:

(a) For every phrase α, there is a noun phrase $\langle\alpha\rangle$, which names it.

(b) There is a phrase *Sat*, requiring two noun phrases to be inserted to make a sentence, such that every sentence of the form $Sat(t\ \langle\alpha\rangle)$ iff $\alpha(t)$

is true, where α is any phrase requiring a noun phrase, t, to be inserted to make a sentence, and parentheses mark insertion.

(c) The following inference principle is truth-preserving:
α iff it is not the case that α
Hence, α and it is not the case that α

The proof that these are jointly inconsistent is analogous to the one given above, and suffices to show that any language satisfying (a)-(c) is inconsistent.

Priest dismisses the possibility of denying (a) and instead focuses on how one might attempt to denying either (b) or (c), considering them in reverse order.

The only reason Priest discusses as plausible for giving up (c) is the existence of truth value gaps. Priest characterises these gaps as 'sentences which are neither true nor false' (2006, p.13). Some care must be taken in interpreting this. Though a number of philosophers have attempted to avoid the paradoxes by denying that the laws of excluded middle and bivalence hold, at least unrestrictedly (famously Kripke (1975) and, more recently, Field (2008)), very few (if any) actually admit straightforward counterexamples to these laws. They don't, for example think there is a sentence α such that $\neg(\alpha \vee \neg\alpha)$; nor do they think there is a sentence, β, such that $\neg(T\langle\beta\rangle \vee F\langle\beta\rangle)$. The reason for this is that, with little more than the De Morgan equivalencies, such sentences lead to contradiction. They do, however, deny the law of excluded middle holds of sentences like the liar, in the sense that they assert that it is neither determinately true nor determinately false, and perhaps it's reasonable to say that, in some sense, those sentences fail to be either true or false (instead, normally, taking a third, 'gap' value), but 'fail' here cannot simply be negation, on pain of contradiction.

So long as the conditional employed by these theories does not simply take the gap value when either its antecedent or consequent are assigned gap, the theory

can invalidate principle (c), since $\alpha \leftrightarrow \neg\alpha$ may be true but $\alpha \wedge \neg\alpha$ fail to be true (presumably taking gap).

Priest suggests two ways in which this might be done (2006, p.13). The first is to take sentences to be truth-bearers, but reject that every sentence is either true or false. The second is to take what sentences generally express, propositions, to be truth-bearers, and to claim that some sentences fail to express a proposition. The essential problems Priest sees for attempts to deny (c) are independent of these competing accounts of truth-bearers, so I will not discuss them further. I will, however, note that the standard terms one must use here to discuss these matters, such as 'sentence', 'statement', 'proposition' and so on are not neutral on the issue of truth-bearers, and genuinely neutral terms are hard to find. So I use these terms more or less interchangeably and take it that partisans in the truth-bearer debate can recast the discussion in their favoured vocabulary without affecting the substance of the issues.

3.2.1 Paracompletism

The best known view which rejects the inference in (c) to avoid paradox is the paracomplete view first described in Krikpe's classic paper *Outline of a Theory of Truth* (1975). The view is still one of the main competitors as a solution to the paradoxes, especially the version refined and advanced by Field (2008).

For these paracompletists, the explanation of the liar's gappiness is that the truth predicate is ineliminable – the sentence fails to be *grounded* in truth-free sentences. Since only grounded sentences can be true or false, and the liar is not grounded, it is assigned gap, and fails to be true or false.

Priest offers a number of criticisms of this sort of view. The first is that certain sentences do not receive the truth-values they ought to. Take a sentence α which is not true at any fixed point. If truth at the fixed point is supposed to model natural language truth (which is the purpose of the construction) , then 'α is not true' ought to be true at the fixed point. But it does not: 'α is not true' receives the gap value at the minimal fixed point.

A second concern Priest raises for these approaches concerns the apparent asymmetry between the liar sentence and the truth teller:

(Liar) The sentence Liar is false
(Truth-teller) The sentence truth-teller is true

Intuitively, in the case of Truth-teller, the sentence can consistently be either true or false, but there seems nothing to determine which it is. This, Priest grants, seems an obvious candidate for a truth-value gap. Liar, on the other hand, seems, on the face of it, to have its truth-value over-determined: we can show by seemingly-plausible reasoning that it is both true and false. This is a more plausible candidate for a truth-value glut (dialetheia) than a gap, according to Priest.

Whether Liar seems plausibly a truth–value glut will depend, presumably, on how plausible one finds the existence of truth-value gluts. Many (perhaps most) philosophers find the existence of such gluts extremely implausible, and so would find it implausible that Liar is an instance of this phenomena. The force of Priest's worry is, I take it, that there is an apparent asymmetry between the two sentences which is not explained by their truth-value assignments (both gap) on the paracomplete theory.

It is unclear how much this objection ought to worry a paracompletists. That there are differences between the two sentences is clear: the most obvious one being the existence of an argument to contradiction from Liar, but no such argument in the case of Truth-teller. Paracompletists can accept this. But the assumption that these differences must result in a difference of truth-value in the sentences seems question-beggingly to depend on the assumption that the argument to contradiction in the case of Liar is sound. The paracompletists central claim is that groundedness is the determinant of whether a sentence receives a genuine truth-value or is assigned gap. Both of Liar and Truth-teller are ungrounded and so it seems quite in keeping with the view that both be treated as truth value gaps.

Priest's next argument against this approach concerns what he calls 'extended paradoxes'. Paracompletists claim that some sentences are truth-value gaps and, on this basis, maintain that though paradoxical sentences like Liar are true

if and only if they are false, since they are neither, we can block the inference to their being both true and false. There is, Priest says, a 'standard ploy' to show that this does not work, which relates to my earlier point about how best to explain the sense in which some sentences fail to be either true or false.

If the paracompletists is correct, one might think, some sentences are neither true nor false and this can be expressed in English using, as has just been done. In particular, for any sentence α which is neither true nor false, 'α is not true' must be true. This does not mean, Priest points out, that 'α is true' is false, because we might allow the negation of a sentence assigned gap to be true. However, the paracompletists ought to accept, Priest suggests, the following:

(*) If α is not true, then 'α is not true' is true

But consider the following sentence, often called the 'strengthened liar'

(SL) The sentence SL is not true

If we suppose this sentence true, then by the T-scheme it is not true. If we suppose it false, then, on the assumption that falsities are not true (which ought to be acceptable to all but certain dialetheists, such as Priest), it is not true and, hence, true. If we assign it gap, as the paracompletists view would seem to demand, then it would seem to be neither true nor false and hence, by (*), we have that (SL) is not true and we have a contradiction. A paracompletist might reject this appearance and take the view that one should be agnostic about whether liars like (SL) are true or false, but assert that they are neither determinately true nor determinately false. In this case, one might reformulate the problem with a sentence saying of itself that it is not determinately true.

Priest grants that it might be objected that this proof assumes that (SL) is either true or false, and hence, assumes excluded middle. But paracompletists do not claim that *every* instance of excluded middle is false and, if they claim (SL) receives gap, and sentences which receive gap are not true, then all we require to establish that (SL) is either true or not true is the disjunctio –introduction rule 'weakening', which is not in dispute.

This problem is widely known, as Priest notes, and the response forced on paracompletists is generally to insist that the notions involved in expressing the

paradox (especially the notion of a truth-value gap) are inexpressible in the language in question. But, according to Priest, to concede this is simply to admit that the theory offered is not a model of natural language, since these notions *are* expressible in English. If this route is taken, the paracompletists has effectively admitted that, were their theory true, it would be inexpressible. This recurring problem which returns for non-dialetheists 'like a bad penny', according to Priest, is the problem of Revenge, discussed in detail in the second chapter of this thesis.

There have been a number of responses to this problem on behalf of paracompletism, the most impressive coming from Field's (2008) book *Saving Truth from Paradox*, whose technically innovative development of Kripke's theory is, as well as attempting to improve on Kripke's conditional, aimed at avoiding this problem. Priest is not convinced, and responds at length to Field's book (Priest, 2010a), arguing that the revenge problem remains (and must remain), for the paracompletist, as intractable as ever. The purpose of this thesis, though, is not to assess the prospects for a paracomplete response to the revenge problem, but to see if dialetheism has this problem too. As should be clear from the preceding discussion, it would be a sore blow to the dialetheist case against paracompletism (and, hence, to the case for dialetheism itself), if dialetheism too suffers from this problem.

3.2.2 Rejecting the T-Scheme

So the attempt to avoid contradiction by invalidating (c) via a rejection of excluded middle seems to face the problem of revenge. The next strategy Priest considers is to avoid contradiction by rejecting (b): in other words, rejecting the Tarski satisfaction scheme. In fact, all Priest's discussion focuses, not on satisfaction, but on truth, and the corresponding T-scheme:

$T\langle\alpha\rangle \leftrightarrow \alpha$

Where α is any closed, non-indexical sentence. The reason he gives is that both principles ought to be treated the same, and discussing truth is both simpler and more in-keeping with the orthodox views he discusses.

Priest begins with the observation that English clearly contains a truth-predicate and that there is a strong presumption that it satisfies the T-scheme. Aside from its intuitive obviousness, one powerful reason to suppose the T-scheme holds is the crucial role that both directions of the biconditional play in allowing truth to do the generalising job for which it is so essential in English.

Suppose Bob utters some sentences, A, B and C and I wish to agree with him. I may simply repeat what Bob said, but this strategy will only carry me so far. Suppose Bob said not just A, B and C, but a great many other things too. It would be, at the very least, cumbersome to have to repeat all the things he said in order to agree with him. Alternatively, suppose I wish to agree with Bob, not on the basis of having heard him utter A, B and C and my believing each of these sentences, but because I believe Bob to be very knowledgeable about the matters A, B and C concern. I may not even have heard Bob's particular utterances, but still wish to agree with him, on account of his superior knowledge. I am in no position now to repeat A, B or C and so, to agree with him, I must say 'What Bob said is true.'

To take a different example, I may know very little of physics, but on the basis of some scientific authority's assertions, want to agree with his endorsement of, say, quantum mechanics. I may be completely unable to understand quantum physics and so be unable to list any of its central claims, but still wish to endorse the theory on the evidence that those who do understand it believe it to be true. It is difficult to see how this could possibly be achieved without my saying something like 'Quantum mechanics is true.'

Some philosophers (known as 'deflationists' or 'redundancy theorists') believe that this generalising job is, in some sense, all there is to the notion of truth. But that it plays this role at all ought to be uncontroversial. However, for truth to play this role, it would seem to be necessary that it obey both of the following in full generality:

(Capture) $\alpha \rightarrow T\langle\alpha\rangle$

(Release) $T\langle\alpha\rangle \rightarrow \alpha$

For, if it fails to obey Capture, then there are things which are the case (that is, we have α), but which fail to be true (we don't have $T\langle\alpha\rangle$). If it fails to obey

release, then there are some truths (we have $T\langle\alpha\rangle$) which are not the case (we don't have α). So, *prima facie*, failure of either capture of release prevents truth from playing the generalising role it so obviously does in English. This would seem good reason for, at least, a strong presumption that the scheme is true.

Priest discusses three reasons the T-scheme might be supposed to fail. The first is the existence of truth-value gaps. The objections already discussed apply as much in this context as in the previous. A further reason Priest gives is that admitting gaps does not straightforwardly falsify the scheme. Suppose we take a sentence, α, assigned gap and argue that it refutes the T-scheme in the following way: since α is assigned gap, α is not true and hence $T\langle\alpha\rangle$ is false. Hence, $\alpha \leftrightarrow T\langle\alpha\rangle$ is false. This reasoning may be resisted, according to Priest, in two ways. Firstly, whether it follows from the fact that α is not true that α is false depends on the truth conditions of the truth predicate, given the existence of truth-value gaps, and the type of negation in play. It may be that, if α is assigned gap, $T\langle\alpha\rangle$ is also assigned gap, not false. And though, he claims, if α is assigned gap, $\neg T\langle\alpha\rangle$ must be true, it needn't be that $F\langle\alpha\rangle$, if the negation is of the 'external' kind which takes a sentence assigned gap to truth.

Secondly, regardless of whether $T\langle\alpha\rangle$ is assigned false or gap, we do not know the truth-value of the biconditional until we have specified the semantics of the conditional involved. For example, if we stipulate it obeys gap in/gap out, then the assignment of gap to α will be enough for an assignment of gap to the biconditional, independently of the value assigned to $T\langle\alpha\rangle$. And if the biconditional is such as to be assigned truth when both sides are assigned the same value, then both α and $T\langle\alpha\rangle$ receiving the same value, whatever that is, will be sufficient for the truth of the instance of the T-scheme. Finally, Priest suggests, if the conditional used is a relevant one (an entailment connective), the truth values of α and $T\langle\alpha\rangle$ are not enough to settle the truth-value of the biconditional.

Priest also provides the following argument, independent of the particular behaviour of the conditional, for why the scheme should hold. The inference from α to $T\langle\alpha\rangle$ is necessarily truth-preserving, since if α is true, so must be

$T\langle\alpha\rangle$. Similarly, if it's true that $T\langle\alpha\rangle$, then it must be that α is true too, so the converse is also necessarily truth-preserving. Since both inferences are clearly also relevant, this, Priest claims, satisfies everyone's informal conditions for a true biconditional.

The first two arguments Priest gives are not themselves demonstrations that one cannot falsify the T-scheme by allowing truth-value gaps. Rather they show that the falsity of the T-scheme is not automatic, once one has accepted truth-value gaps, and will depend, among other things, on the behaviour of the conditional invoked. The immediately preceding argument, however, is a direct argument against this approach. Effectively, Priest's claim is that since the inference from α to $T\langle\alpha\rangle$ (and *vice versa*) is necessarily truth-preserving, α entails $T\langle\alpha\rangle$ (and *vice versa*) and that this is enough to make a biconditional true on anyone's 'informal' conditions. Defenders of the view under discussion, then, are committed to the T-scheme anyway.

It is unclear what 'informal' conditions are in this case, as opposed to formal ones, or what the relationship between these conditions ought to be. Perhaps it is an appeal to the intuitive obviousness of the inference from α to $T\langle\alpha\rangle$ (and *vice versa*). In any case, it seems to me that a paracompletist inclined to go the route Priest is discussing may respond by denying that entailment is a matter of necessary truth-preservation, but rather a matter of necessary preservation of designated value (non-false value).

If they endorse the view that there are truth-value gaps, and that this violates the T-scheme because when a sentence α is assigned gap, $T\langle\alpha\rangle$ is assigned false, then α has been assigned a designated value, but $T\langle\alpha\rangle$, not. Hence, the inference from α to $T\langle\alpha\rangle$ does not necessarily preserve designated value and is invalid. Presumably such a paracompletist would wish to defend this as an account of informal validity and would resist the premise in Priest's argument above that α entails $T\langle\alpha\rangle$. This is not to say that such a view is attractive, but merely that it is not shown to be false by Priest's argument.

Finally, Priest gives a third reason that attempting to invoke truth-value gaps to falsify the T-scheme will not avoid contradiction. If we take a strengthened liar sentence, α, equivalent to 'α is not true', paracompletists are committed to this

being assigned gap, and hence, plausibly, to the claim that α is not true, and hence to α. In other words, Priest says, α has a truth value and the relevant instance of the t-scheme holds. This is the familiar revenge objection from the previous section, which again seems to be bearing significant argumentative weight in Priest's case against the paracompletist.

3.2.3 Hierarchical Approaches

The next class of solutions Priest discusses (2006, p.19) are hierarchical solutions, mainly developed from the work of Tarski. According to the first sort of hierarchical theory, the paradoxes result from the mistaken assumption that languages like English are unified, semantically closed languages. In fact, so the theory goes, English is a hierarchy of languages L_i, each with a truth-predicate T_i, which applies only to sentences in the level below in the hierarchy. The effect of this is to rule-out as ill-formed putative liar sentences, since they attempt to predicate truth of a sentence in their own language. It should be noted that, although such theories make essential use of Tarski's formal machinery, he himself did not endorse this formal picture as a model of natural language, which he took to meet the conditions specified in Section 3.2, and thus to be incoherent (Tarski, 1936).

Priest claims a number of problems afflict this sort of view. The first is the apparent lack of evidence for the claim that English is a hierarchy of languages in this way, or that all uses of a truth-predicate in English are ambiguous between infinitely many truth predicates, each indexed to a distinct language in the hierarchy. It is neither intuitively obvious that this is the case, nor is there any evidence, linguistic or grammatical that this is the case. This, he says, is sufficient to make the view unsatisfactory. This seems a little quick. Dialetheism is not intuitively true, and some inferences which dialetheism is forced to claim are invalid, are intuitively valid. There is also very little evidence, to my knowledge, that people are wont to assent to contradictions, and a lot of evidence that they are wont to reason by disjunctive syllogism. Clearly this is not, on its own, reason to rule dialetheism unsatisfactory, though it seems good reason to be cautious of the view. Similarly, the absence of evidence of the kind

Priest mentions for the hierarchical view, warrants caution, but doesn't seem quite a refutation.

Priest does, however, discuss more powerful arguments against the view. One concerns a sentence which I adapt here as 'Every sentence in Section 3.2.3 of this thesis is true.' This is, he suggests, a perfectly clear sentence of English which, on the assumption that at least one other sentence in this section is false, is false. Yet the sentence cannot occur within the hierarchy, because it predicates truth of itself. Since the sentence occurs in English, but not in the hierarchy, Priest says, the hierarchy is not English. Priest attributes this, and related objections, to Kripke (1975) and his general observation that it is not intrinsic syntactic and semantic features of a sentence which render it paradoxical but also other, external factors. So any theory which focuses solely on the former to rule-out paradoxical sentences will also rule-out other, ordinary sentences and hence will be expressively weaker than English.

A further objection to the view is characterised by Priest as follows. We call the language with the lowest ordinal at which a particular sentence occurs that sentence's 'rank'. The sentences which are true in the hierarchy are those which are true at their rank. We then construct a liar-type sentence which says, intuitively, 'this sentence is not true at its rank'

(a) $\neg T_{rk(\alpha)}<\alpha>$

If this sentence occurs in the hierarchy, it has some rank, i. By the T-scheme for rank i,

$T_i<\alpha> \leftrightarrow \neg T_{rk(\alpha)}<\alpha>$

Substituting i for $rk(\alpha)$ delivers contradiction. The only response to this, according to Priest, is for the proponent of the view to deny that this is a sentence of the hierarchy, and that quantifying over the indices of truth-predicates, which is required to define the notion of rank, is inexpressible in the hierarchy. But since these claims are expressible in English, Priest claims, this is simply to admit again that the hierarchy is not English. Indeed, to even characterise the hierarchical view, one must assert the existence for each i of a truth-predicate, T_i, which, since disallowed by the hierarchical view makes the view, if true, inexpressible. This, again, is the revenge problem.

A second class of hierarchical view discussed by Priest takes English to be a single language which enjoys a single, univocal truth-predicate, but takes interpretations of it to form a hierarchy (Priest refers here to the views of Gupta (1982) and Herzberger (1982)). The interpretations in the hierarchy differ only in their assignment of the extension of the truth predicate. According to Priest's characterisation, given an arbitrary extension assigned *T* at level 0, the extension at ordinal level *i*, which we call E_i, is the set of sentences, α, such that for some ordinal $j < i$, and all ordinals, *k*, such that $j \leq k < i$, α holds in the interpretation at *k*. At a certain height, the hierarchy stabilises. Some sentences enter the extension of the truth-predicate and stay there for the remainder of the hierarchy and we call these 'stably true'. If the negation of a sentence is stably true, we call it 'stably false'. We call an interpretation 'stabilised' if all stable sentences have their ultimate value.

According to Priest, the unstabilised interpretations are, in some sense, unimportant, since it's the stabilised interpretations which give the important properties of sentences. Given a sentence α, the T-scheme for α is guaranteed to be true at a stabilised interpretation if and only if α is stable. Paradoxical sentences such as the liar are unstable and so the T-scheme may fail for those at stabilised interpretations.

Priest offers a number of criticisms of this view. The first is, he claims, it is unclear what it is supposed to show: that is, the relationship between this formal structure and (the semantics of) English is unclear. It is unclear, according to Priest, what it means to say that English has a whole hierarchy of interpretations.

A second objection is that the hierarchy must, in some sense, characterise the meaning of the language under consideration, and meaning is something grasped by the users of that language. So, if this theory were a characterisation of English, speakers of English would be able to grasp the notions it employs in characterising meaning. The notions this view presupposes are grasped by ordinary speakers of English include, according to Priest, transfinite ordinal, definition by transfinite induction and others. The implausibility of this, he says, need not be laboured.

Another objection Priest gives is as follows. Truth, he says, is the aim of a number of cognitive processes, such as asserting, theorising, and so on. The extension of the truth-predicate on the view under consideration varies from interpretation to interpretation. Which, Priest asks, is the target set? He considers a few options. The first, which is unlikely to be found attractive by anyone endorsing the view, is that we take the process of ascending the ordinal ranking of interpretations to be temporal. The choice of temporal unit would, Priest points out, be arbitrary. Worse, it would lead to the result that the liar might be assertible, say, on Mondays, Wednesdays and Fridays, but not on Tuesdays, Thursdays and Saturdays, for example. Another alternative is to take the true to be the union of all the extensions of the truth predicate (or perhaps just those in stable interpretations). But this would be inconsistent, and so would fail to avoid dialetheism. A final option Priest considers, which he takes to be the most plausible suggestion, and the one most likely to be endorsed by proponents of this view, is to take the intersection of all these sets (the set of stable truths) to be the target. The effect of this, however, would be to trifurcate sentences into the true (stably true), the false (stably false) and the rest (neither stably true nor stably false). But this is just the gap view already discussed.

In particular, the revenge problem which Priest claims affects the gap view now returns for the present view:

(β) The sentence *β* is not stably true

Since classical logic holds at each stage in the hierarchy, *β* is either stably true or not. If it is stably true, it is true at every stabilised interpretation in the hierarchy and, hence, by the T-scheme (which, again, holds at every stabilised interpretation), '*β* is not stably true' is stably true. Hence we must assert something which says it can't be asserted. So we have shown, and can therefore assert, that *β* is not stably true. Hence we can assert something which says it can't be asserted. A more precise characterisation of this argument is given in Priest (1987).

One response Priest discusses to this problem comes from Gupta (1982). The idea is that the extension of 'is stably true' varies up the hierarchy. The obvious way to do this, Priest says, is to take the extension of the predicate at *i* to be

the set of things which, at that level, appear to be stably true. The result of this, according to Priest, is that the truth-predicate and the stably-true-predicate become coextensional, and so the latter, vacuous. Priest points out that, even were a characterisation of the stable truth predicate given, it is unclear how this would help with the relevant liar-paradoxical sentences. The suggestion under consideration is that utterances about stable truth themselves may be unstable; but the argument from extended liar sentences does not obviously assume otherwise. Thus, according to Priest, the only real solution available is to deny that stable truth is expressible in the language and, since it is expressible in English, to deny that the language in question is English.

3.3 Set-Theoretic Paradoxes

3.3.1 The Inconsistency of Naïve Set Theory

I now turn to Priest's account of the set-theoretic paradoxes and his defence of a dialetheic solution to them. The situation with these paradoxes is different from that of the semantic paradoxes in a number of ways, according to Priest (2006, p28). The first is that there is a more or less clear account of what the naïve notion of set is and, moreover, a general consensus about what is wrong with it.

Another difference is that semantics, even the formalised kind, has never been a mathematical theory in the sense of being studied for its interesting mathematical properties, or its ability to produce interesting mathematical results. This, he thinks, means that in debates about naïve semantics, there is no well-developed, generally agreed mathematical practice to which we can appeal to settle our disputes.

A further, important difference is that the formalisation of semantics came after, and at least in part, as a result of, the semantic paradoxes. The result being that, for Priest, the formalisation is 'far too self-conscious to be of much help'. The point, I take it, being that rather than simply being an attempt to

characterise our naïve understanding of various semantic notions, people like Tarski (and those who followed him) were well aware of the threat of inconsistency and attempted to avoid it, perhaps in an *ad hoc* way, and at the cost of losing the naiveté of the account. This is clearly true of Tarski, at least, who, as I have mentioned, was explicit about his theory *not* being an account of the naïve semantics of natural language, which he took to be obviously inconsistent.

The situation with set theory differs importantly from semantics, Priest thinks, in large part because the development of set theory began before any awareness of the set-theoretic paradoxes. Even after early set-theorists became aware of certain of the paradoxes (for example, Cantor's seemingly early recognition of the paradox that bears his name; though it seems he did not think of it as a paradox), at least some developments of set-theory, such as Frege's, were characterised without knowledge of the paradoxes. So Frege's characterisation was not, in Priest's words, 'deformed' by attempts to avoid paradoxes, or the *ad hoc* refinements of our naïve understanding which such attempts often seem to involve. According to Priest, this gives us good reason to suppose Frege's theory an accurate account of our naïve understanding of set. The two pillars of Frege's theory are:

(*Abs*) $\exists y \forall x\ (x \in y \leftrightarrow \beta)$

(*Ext*) $\forall x(x \in z \leftrightarrow x \in y) \rightarrow z = y$

With the restriction (added by Priest, though not absolutely required) that β is any formula not containing y free. Priest points out that this is not exactly Frege's formulation of the theory, which is second order with '$\in$' defined, but it is close enough for our purposes. The second principle, *Ext*, stipulates that sets are extensional entities, identified by their members (as opposed, perhaps, to what some philosophers take to be intensional entities, like properties which might be such that two distinct properties have the same extension). The first, and perhaps most important, principle stipulates that every condition has a set as its extension: in fact, on this view, we may take a set just to be the extension of an arbitrary condition.

As was famously pointed out by Russell, the conception of set given by *Abs* and *Ext* is inconsistent:

$\exists y \forall x (x \in y \leftrightarrow x \notin x)$
$\exists y (y \in y \leftrightarrow y \notin y)$
$\exists y (y \in y \wedge y \notin y)$

The first step being the instance of *Abs* giving the existence of a set of all non-self-membered sets, the second, which follows from the first, is that this set is a member of itself if and only if it is not and the final is an application of the *reductio* rule, equivalent to excluded middle.

A further argument Priest gives against the strategy is that not all of the set-theoretic paradoxes seem to involve appeal to excluded middle. One is the Burali-Forti paradox, since, he says, a direct argument can be given that the set of all ordinals is an ordinal and an independent argument given that this is not the case. The contradiction that it both is and is not a member of itself is therefore delivered by an instance of conjunction introduction from the conclusions of these distinct arguments, without the need for appeal to an intermediate conclusion according to which the set is an ordinal if and only if it is not.

Another example which Priest discusses in more detail (Priest, p.29) is Miraminoff's paradox, which arises when we consider whether the set of all well-founded sets is well-founded. We define a *regress* from z to be a function, f, from the natural numbers such that $f(0) = z$, and for all n, $f(n) = \emptyset$ or $f(n + 1) \in f(n)$. We call a regress *bounded* if, for some n, $f(n) = \emptyset$. Let W be the set of all well-founded sets: the set of all sets z such that every regress from z is bounded. Let f be any regress from W. If $f(0) = \emptyset$ the regress is bounded. Suppose $f(1) \in W$. Let g be the function such that $g(n) = f(n + 1)$. g is a regress from $g(0) = f(1) \in W$. So g is bounded and, hence, f is bounded. So all regresses from W are bounded and $W \in W$. But consider the function h such that, for all n, $h(n) = W$. h is an unbounded regress from W and so $W \notin W$.

Informally, we can construct a regress from a set, expressing the fact that the set is a member of another set, this of a further set and so on. If all the regresses we can construct from a set are finite, i.e. they all eventually

'bottom-out' in a final member (in the case of pure set theory, this will always be the empty set) , the set is well-founded. If there is at least one chain which is infinite, i.e. there is no final member of the chain, the set is non-well-founded. The simplest way to illustrate the latter is to consider a set, say, α, which is a member of itself. Since α heads the regress $\alpha \in \alpha \in \alpha$… and so on indefinitely, α is non-well-founded; the same goes, obviously, for any self-membered set.

Now, consider the set *W* of all well-founded sets: this set must be well-founded for, if none of its members heads an unbounded regress, then any regress headed by *W* must be bounded, since any such regress would have to proceed through one of *W*'s members. But now we can also show that *W* is non-well-founded, since it is self-membered and thus heads the regress $W \in W \in W$…

This, according to Priest, establishes the inconsistency of the naïve notion of set even if one denies excluded middle, since the proof makes no use of the latter. There are three responses available here: one might deny that arguments of this kind establish the inconsistency of naïve set theory; or one might accept the inconsistency of naïve set theory and abandon it on this basis; or one might accept the inconsistency of naïve set theory, but take this only to show that one ought to reason about sets paraconsistently.

The second of these options is by far the most popular. Indeed, it is almost universally endorsed by both mathematicians and philosophers, who generally replace the naïve characterisation of set with the one specified by the standard Zermelo-Fraenkel axiomatisation of set theory (perhaps with the addition of the axiom of choice).

Priest's preference is, of course, option three: one should keep the naïve account of set and accept its inconsistency. His arguments for this approach, which I will describe presently, are generally concerned with attacking the orthodox position, option two.

This, however, leaves unscathed option one, the defender of which might happily endorse Priest's arguments against orthodoxy (and in favour of the naïve account) whilst demurring from the conclusion that dialetheism be endorsed. This view is defended by Alan Weir (in, for example, his (1998)) who attempts to avoid the paradoxical arguments for the inconsistency of naïve set theory by

restricting the structural rules of the underlying logic (specifically, the structural cut-elimination rule). Since my focus here is dialetheism, I will not discuss this view in any depth.

No dialetheists, so far as I am aware, have attacked Weir's views on set theory directly yet; but this will have to be done if Priest's arguments are to persuade us, not just of naïve set theory, but of dialetheism. There are costs, of course, to Weir's view (the most obvious being that there is a strong presumption that the structural rules he asks us to abandon are correct) but whether these costs are significant enough to make a dialetheist approach to naïve set theory the most plausible remains to be seen. One argument a dialetheist such as Priest may wish to give here is that, since Weir's view is consistent, it cannot give a revenge-free treatment of paradoxes such as the liar. A discussion of whether this is correct is outwith the scope of this thesis, but if this is the best response to Weir's view on behalf of dialetheism, it makes all the more clear the significance of the revenge problem in making a case for inconsistency.

3.3.2 Priest Against Set-Theoretic Orthodoxy

3.3.2.1 Lack of Motivation for the Cumulative Hierarchy

The target of the next part of Priest's defence of dialetheism is the orthodox conception of set given by the cumulative hierarchy (normally as specified in Zermelo-Fraenkel (ZF) set theory), against which he defends the naïve conception of set given by (*Abs*) and (*Ext*).

Priest begins by claiming that the burden of proof lies with his opponent. If they grant, as he suggests most set theorists do, that (*Abs*) and (*Ext*) characterise the intuitive notion of set, then they must offer an account of what they take to be the correct account of set, and an explanation of why (*Abs*), in particular, fails.

The cumulative hierarchy is obtained by beginning with (at least in the case of set theory without urelements) the empty set and taking its power set (the set of all its subsets), then taking the power set of this, and so on. This is iterated into the transfinite and each level in the hierarchy is indexed by an ordinal

number which gives its rank. Whenever a limit ordinal is reached (one such there are ordinals smaller than it and, for every such, there is a further ordinal between it and the limit), we progress by taking the set of everything which has come before. We stipulate that the only true instances of (*Abs*) are those which hold in this hierarchy and thus block the paradoxes, since the sets which generate them do not occur in the hierarchy.

The first reason Priest gives against this view is that its intelligibility depends on a prior grasp of the notion of ordinal for, unless we know 'how long' the construction goes on for, we haven't grasped the nature of the hierarchy. But ordinals are set-theoretic entities or, at least, they must be characterised set-theoretically.

If we endorse the naïve characterisation of set, according to Priest, we can characterise the ordinals and thereby define the cumulative hierarchy. But if the cumulative hierarchy relies on the ordinal numbers, the standard view must presuppose some other notion of set by which we grasp these. On the possibility of a theory such as the aforementioned *ZF* specifying the operative notion of ordinal, Priest says that this would introduce vicious circularity into the theory, since the rational for *ZF* presupposes the cumulative hierarchy.

Another problem with supposing that only certain instances of (*Abs*) hold, according to Priest, is that there would appear to be no obvious explanation for why (*Abs*) seemed plausible to us in the first place. From the perspective according to which only certain instances of (*Abs*) hold, he thinks, (*Abs*) does not seem plausible. On Priest's own view, the plausibility of (*Abs*) is explained by its truth and so, he claims, his view is explanatorily superior.

This doesn't seem quite fair. Our pre-theoretical believes about sets and related logical notions lead not just to inconsistency, but to triviality. If we wish to avoid this, as Priest himself does, some of our pre-theoretical beliefs must be given up. In his case, we must endorse a paraconsistent logic which invalidates, among other things, disjunctive syllogism. We certainly have a pre-theoretical belief that disjunctive syllogism is valid, but since it's not on Priest's view, we might ask why it seemed valid in the first place, if so many of our notions turned out to be inconsistent. Someone who endorses, say, classical logic, can explain why disjunctive syllogism seems valid by appealing to its actual validity.

Priest's answer (defended in Chapter 8 of his (2006)) is that, in most ordinary contexts, disjunctive syllogism is truth preserving. It's only when we start thinking about totalities which result in paradoxes that we discover dialetheia and hence discover that disjunctive syllogism can't be valid. But a defender of the cumulative hierarchy may wish to say much the same thing about (*Abs*): a great many of its instances are true, and it is only when we consider paradoxical totalities that we realise we must give up some of its instances.

The paradoxes demonstrate that anyone who wishes to avoid trivialism must give up some or other of our intuitive, pre-theoretic beliefs about sets, or logic, or semantics and from the perspective of the resulting, post-paradoxical view, those notions are apt to seem implausible in retrospect. It's neither obvious that this is a genuine problem, nor that it is avoidable. At the least, further argument is needed that dialetheism has a distinct advantage on this score.

Another objection Priest gives is that the claim that the only sets are those in the cumulative hierarchy lacks independent justification. This, he claims, is sufficient to make a view unsatisfactory. This may also be a little quick: whether this makes a view unsatisfactory may depend on whether the other views available can do better. It may be that certain theoretical decisions must be made simply because they avoid the paradoxes, or perhaps because they avoid triviality. A theory which avoids this is clearly, *ceteris paribus*, at an advantage, but it doesn't seem right that a theory be automatically disqualified because certain of its features are motivated solely by considerations deriving from the paradoxes.

Priest cites as a standard attempt to motivate the hierarchy, Devlin (1980):

> Fundamental to set theory is the concept of being able to regard any collection of objects as a single entity. But before we can form a collection of objects, those objects must first be "available" to us…Before we can build sets of objects, we must have the set of objects out of which to build these sets. The crucial word here is, of course, 'build'. Naturally we are not thinking of *actually building* sets in any sense, but our set theory should reflect this idea.

Priest criticises this rational in a number of ways. Firstly, it employs an inappropriate temporal metaphor, suggesting that sets come into existence after

their members, which is not true (though Devlin does admit that sets are not actually built ‘in any sense’). An alternative metaphor which would work equally well, says Priest, is of starting with the biggest set and building down ‘as one might build a chain suspended from a rafter’ (Priest, 2006, p.32), which would allow for non-well-foundedness.

The only non-metaphorical sense to be made of the building metaphor, according to Priest, is to take the claim to be something like ‘a set is conceptually prior to its members’. Though it’s unclear, Priest thinks, that exactly what this means, he points out that it doesn’t seem to rule out non-well-foundedness, since there’s no reason this dependency should always bottom-out in a well-founded way.

Priest discusses one attempt to show that there is no indefinite regress here, due to Mayberry (1977). According to Mayberry, the determinacy of any set depends on the determinacy of its members. Thus, on this view, any argument for the determinacy of a set must be given via arguments for the determinacy of its members. So, to argue for the determinacy of non-well-founded sets would require non-well-founded arguments; but, says Mayberry, ‘it is obvious that no argument whose premises proceed in a circle or a regress to infinity can be valid’ (Mayberry, 1977, p.31).

This, according to Priest, does not work. First, it is not true that regressive arguments are invalid, since α, because α, because α… is a ‘paradigm of validity’. Though Priest doesn’t discuss the case, neither is it true that circular arguments cannot be valid since $\alpha \vDash \alpha$ is clearly valid.

As Priest points out, neither must an argument that some set x is determinate have ‘y is determinate’ as a premise for each $y \in x$. There, for example, be a single premise (perhaps (*Abs*)), which entails the determinacy of x and all its members.

Finally, he notes, the argument depends on the assumption that, if one cannot show by argument that something is determinate, it follows that the thing is indeterminate. Which is fallacious.

3.3.2.2 Category Theory

A further consideration of Priest's against the view that the cumulative hierarchy exhausts the universe of sets is its inadequacy in category theory. Category theory is a highly abstract theory of structure whose objects are certain algebraic entities, the categories, and the structure preserving relations (morphisms) which hold between them.

One of the interests of category theorists is in studying the structural features of entities described by other highly abstract mathematical theories; so they are interested, for instance, in the category of sets and the category of groups. But these are the very sorts of totalities disallowed by defenders of the cumulative hierarchy. Worse, according to Priest, category theorists want to carry out operations on these categories. For example, from the category of sets, V, and the category of groups *Grp*, they might wish to create the functor category, V^{Grp}, of all functors from *Grp* to *V*. But these constructions, says Priest, are not possible in the cumulative hierarchy.

Priest considers, and rejects, some suggestions for avoiding this problem. The first is to take problematically large categories (such as the sets or the groups) to be proper classes, where these are subcollections of the hierarchy (all their members occur in the hierarchy), but which cannot be members of any other collection (and, *a fortiori*, don't occur in the hierarchy themselves).

The first problem Priest suggests for proper classes is that they are 'a masquerade' (Priest, 2006, p.34). He claims that 'proper class' is an evasive renaming of certain sets which don't occur in the hierarchy and, if we have to admit that there are sets outside the cumulative hierarchy, this demonstrates the incompleteness of the hierarchy as a picture of the sets. Depending on how this problem is put, a defender of the hierarchy may find it question-begging: to defend the existence of proper classes is only to accept the existence of sets outside the hierarchy if it is assumed that proper classes are sets, which is precisely what proponents of proper classes deny.

However, Priest goes on to say that the insistence that proper classes cannot be members of other collections cannot be given independent motivation. There is

no reason, he claims, why such entities, if they are determinate entities, cannot be members of other collections, such as their singletons. Another example, which seems even more implausible than their inability to be members of a singleton is this: as much as any other kind of mathematical entity, we are able to talk about and use proper classes (using locutions such as, for example, 'the proper classes'), even describe them axiomatically, in theories such as Neumann-Bernays-Gödel set theory, yet we cannot in any sense countenance as a totality 'the proper classes', since it can be neither a set nor a proper class. So we have a type of mathematical entity, all of whose tokens are perfectly determinate objects and seem every bit as much a totality of entities as any other, yet they are somehow precluded from 'coming together' to form a collection.

Some independent motivation must be given for this restriction, or there is considerable justice in Priest's charge that this is simply an *ad hoc* renaming of what are really sets, just to avoid the contradictions generated by the paradoxes (or, perhaps more generally, to avoid revising our logic to endorse naïve set theory without triviality).

Even if this can be done, Priest points out, since proper classes cannot be members of other totalities, we cannot carry out operations on them, for example by forming functor categories of them, such as the one given above. So whilst, properly motivated, the invocation of proper classes might be hoped to give us the existence of the totalities studied by category theorists, it does not make sense of their practice of carrying out operations on those totalities.

In response to this, a defender of the hierarchy might suggest that proper classes can be members of *some* totalities, perhaps 'hyper-classes', but not of other proper classes or of sets. The proper classes, therefore, would form a hyper-class. This response invites two replies: firstly, the fact that we can ask exactly the same questions of hyper-classes as we did of proper classes seems to invite a vicious regress: we will be forced to invoke super-duper-hyper-classes, which will be members of even higher order classes, and so on. We may then wish to form the totality of all the totalities which occur in this hierarchy of classes, and so be back in exactly the same difficulty which was supposed to be solved by the introduction of proper classes in the first place. Secondly, allowing for proper

classes to be members of other totalities makes the insistence that they not be members of sets even more difficult to motivate. The reason proper classes were not allowed to be members of sets was that proper classes are not the sort of entities to be members of other mathematical totalities. But if they can be members of hyper-classes, this is no longer the case, so why can't they be members of sets? These and related objections are made by a number of other philosophers, such as Boolos (1998), Lewis (1991) and Weir (1998).

The second solution Priest considers to this problem is, rather than invoking proper classes, to represent the claims of category theory as being about initial segments of the hierarchy, rather than its entirety. The first way of doing this Priest outlines is to suppose the existence of some inaccessible cardinal κ and, since the sets less than κ have certain useful closure properties, interpret the claims of category concerning the sets as being about the "sub-universe" of sets less than κ. The problem with this strategy, according to Priest, is that, though it produces a model for category theory, it's not the intended one. The point of category theory is to describe the structural isomorphisms holding between all algebraic structures of the relevant kind. But on this view, category theory does not describe the sets, it describes the sets less than κ.

A related, alternative proposal is to suppose the existence of arbitrarily many inaccessible cardinals, each of which provides a 'mini-universe' in which category theory may be interpreted. There are two reasons, according to Priest, that this proposal fails. The first is that the claims of category theorists are still not being interpreted as being about *all* the sets: all of their claims are interpreted as being about one of the mini-universes in the hierarchy. The second is that it suffers from familiar expressive limitations: we want to make claims which cross levels in the hierarchy, but cannot, on this proposal, succeed in doing so. For example, we may wish to claim that some group bears some relation, *R*, to every other group whatever (i.e. at every level in the hierarchy). But on the present proposal, the best that can be done is to interpret this as claiming that, for any mini-universe, there is a group in that universe which bears *R* to all other groups *in that universe*. The claim that the group bears *R* to every group can't, Priest claims, be expressed.

3.3.2.3 Logic

Priest also thinks logic, in particular logical validity, provides further reason to doubt that the cumulative hierarchy exhausts the sets (2006, p.36-37). Consider a first-order language with validity defined as follows:

$\Sigma \vDash \alpha$ iff in every interpretation in which all the members of Σ are true, α is true

Interpretations, understood model-theoretically, are set-theoretic entities whose domains are arbitrary sets. So the quantifiers in the definition above, Priest suggests, appear to range over all the sets. Definitions of the above kind are given in some language (in this case, mathematized English), which presumably has semantics. But, according to Priest, if we restrict ourselves to the cumulative hierarchy, no semantics for the language can be given, for an interpretation is an ordered pair $<D, I>$, where D is the interpretation's domain, but since D in this case is not a set, this interpretation is not available.

As Priest notes, similar responses can be made to this problem as were considered in the previous section, but since similar objections can be raised to these responses in the present context, Priest thinks they fail for the same reasons as before.

This argument strikes me as powerful, and an instance of the more general problem, sometimes called the problem of 'absolute generality' (Rayo and Uzquaino, (2006)), of making sense of expressions (such as 'the sets', 'the ordinals' or even 'everything') which appear to concern totalities normally thought too large to be sets. There are a number of responses to this problem, all of which seem to face serious problems (Rayo and Uzquaino, (2006) contains defences of a number of approaches to the issue). If, however, one is able to endorse naïve set theory (either inconsistently, as Priest does, or consistently, as does, for example, Weir (1998)) absolute generality presents no particular problem. If the central claim of this thesis is correct, and dialetheism cannot be motivated by appeal to revenge arguments, this is one promising avenue for the dialetheist to philosophically motivate their view.

3.3 Paradoxes of Inclosure

A final argument I will consider for a Priest's version of dialetheism concerns the apparent structural isomorphism between the semantic and set-theoretic paradoxes. Though this was not discussed in detail in *In Contradiction* (the reader is referred, on p.263, to his (2002)), in later work (particularly his (1994) and (2002)) Priest has developed a precise account of the sense in which he takes the paradoxes of self-reference to be related: they are all paradoxes of *inclosure*. If this account of the underlying structure of the paradoxes is correct, then combined with the plausible assumption that like problems require like solutions (the Principle of Uniform Solution), this seems to show that the semantic and set-theoretic paradoxes require unified treatment. One significant upshot of this is that it would make Priest's revenge arguments for dialetheism about the semantic paradoxes an indirect argument for dialetheism about the set-theoretic paradoxes. Another is that it puts pressure on those who wish to endorse the classical picture of set theory given by the cumulative hierarchy whilst endorsing a non-classical solution to the semantic paradoxes (this is the view taken by Field (2008), for one, and Beall (2009) for another). If Priest is correct, these views are problematically bifurcated, and their proponents would seem to face a choice between a more thoroughgoing non-classicality, extending to set-theory, or a classical solution to the semantic paradoxes.

Priest first gives an account of what he would later call the 'Inclosure Schema' in his (1994), where he explains the roots of the account in Russell's work. He settles on the following conditions as characterising the structure of the paradoxes (2002, p.134):

(1) $\Omega = \{y; \varphi(y)\}$ exists, and $\psi(\Omega)$
(2) if x is a subset of Ω such that $\psi(x)$: (a) $\delta(x) \notin x$
(b) $\delta(x) \in \Omega$

Where φ and ψ are properties, δ is some function and Ω is a class. Priest calls δ the 'diagonaliser' function, since, paradigmatically, it will be defined by diagonalisation paradigmatically (though, so long as it is defined systematically such that the result of applying it to a set does not deliver a member of that set, diagonalisation isn't essential). Condition (1) Priest calls 'Existence', for the

obvious reason that it stipulates the existence of a set, Ω, of all those things of which φ holds and that ψ hold of Ω. Conditions (a) and (b) are called, respectively, 'transcendence' and 'closure', since if (a) is satisfied, δ transcends the boundary of x and, if (b) is satisfied, Ω is closed under δ. These conditions generate contradiction, since applying (a) and (b) to Ω gives us $\delta(\Omega) \notin \Omega$ and $\delta(\Omega) \in \Omega$, respectively.

This is a generalised version of an earlier schema he calls 'Russell's Schema' (2002, p.129) which lacks the stipulation in the schema above that δ only need operate on the members of Ω's powerset of which ψ holds. Though the paradoxes of absolute infinity (such as Cantor's Paradox) fit Russell's Schema, the generalisation which gives the Inclosure Schema is necessary to capture the paradoxes of definability. For those paradoxes fitting Russell's original schema, we simply take ψ to be some universal property, such as $\lambda x x = x$, and so take δ to operate on the entirety of the powerset of Ω.

We can illustrate how the paradoxes in question are supposed to fit this schema by giving a couple of examples. First, Russell's paradox, as Priest presents it in his (2002, pp. 129-133): we take the property $\varphi(y)$ to be $\varphi \in V$ (where V is the universal set), the property $\psi(x)$, as above, to be the universal property $\lambda x x = x$, the set Ω to be V and $\delta(x)$ to be the function which takes any x to the set of all its members which are not members of themselves ($\{y \in x; y \notin y\}$), which we label ρ_x.

The case of interest is the one in which we take x to be V. We obtain transcendence since, if $\rho_v \in \rho_v$, then $\rho_v \notin \rho_v$ and, hence, $\rho_v \notin \rho_v$. But if $\rho_v \notin \rho_v$, then we have $\rho_v \in \rho_v$ and, hence, we have closure and the contradiction $\rho_v \notin \rho_v$ and $\rho_v \in \rho_v$.

Priest also, explains (1994, p.30) how the familiar Liar Paradox fits the Inclosure Schema. We take the property $\varphi(y)$ to be the property 'y is true' (and so Ω is the set of all true sentences), the property $\psi(x)$ to be the property 'x is definable', and δ to be the function σ, definable by diagonalisation, such that if α is a definable set, $\sigma(\alpha) = \alpha$ where $\alpha = <\alpha \notin \alpha>$ (i.e. α says 'α is not in α').

If α is definable and a subset of Ω, then $\sigma(\alpha) \in \alpha$ and so $<\alpha \notin \alpha> \in \alpha$ (since $\sigma(\alpha) = <\alpha \notin \alpha>$). Therefore $<\alpha \notin \alpha> \in \Omega$ and, by the T-scheme, $\alpha \notin \alpha$ and so (since $\sigma(\alpha)$

= α), $\sigma(\alpha) \notin \alpha$. But if from $\sigma(\alpha) \in \alpha$, it follows that $\sigma(\alpha) \notin \alpha$, then we have $\sigma(\alpha) \notin \alpha$ and so we have transcendence. But, if we have $\sigma(\alpha) \notin \alpha$, then we also have $\alpha \notin \alpha$ and, by the T-scheme, $<\alpha \notin \alpha> \in \Omega$, which means that we have $\sigma(\alpha) \in \Omega$ (again, since $\sigma(\alpha) = <\alpha \notin \alpha>$) and closure, too, is satisfied. The liar sentence here is $\sigma(\Omega)$ and the contradiction is $\sigma(\Omega) \in \Omega$ and $\sigma(\Omega) \notin \Omega$.

An interesting, and perhaps problematic, case is Curry's paradox. Priest points out (1994, p. 33) that for many of the paradoxes he discusses, we can replace $\neg\alpha$ uniformly with $\alpha \rightarrow \perp$ (where '$\perp$' entails everything) and generate a 'Curry' version of the given paradox. Whether these paradoxes are of the Inclosure variety depends, according to Priest, on the meaning of '$\rightarrow$'. If the conditional is material, then, on his view, '$\alpha \rightarrow \perp$' is equivalent to '$\neg\alpha$' and so the Curried versions of the paradoxes are not distinct from the originals and thereby fit the Inclosure Schema just if the originals do. If the conditional is something else, perhaps a strict conditional, then, Priest claims, the Curried versions of the paradoxes are not Inclosure paradoxes.

On the one hand, it is important for Priest that this is so, since, if those paradoxes fit the Inclosure schema, that would seem to demand that they receive a dialetheic treatment. But in the case of a strict conditional, this generates not just inconsistency, but triviality. On the other hand, the Curry paradox is clearly of-a-piece with paradoxes like the Liar, at least in the sense that it is a paradox of self-reference. So if the Curry paradox does not fit Inclosure, one might think the Inclosure Schema can no longer be thought of as giving an account of the structure common to the paradoxes of self-reference. Were this the case, the argument that all paradoxes fitting Inclosure must be given like treatment might be thought to lose some of its bite: if it doesn't capture that which underlies all self-referential paradoxes, why think that what it does capture is important enough to justify the demand for similar theoretical treatment in our accounts of those paradoxes?

A number of problems, some concerning Curry's paradox and some not, have been suggested for the uses to which Priest puts the Inclosure Schema (for example, Grattan-Guinness (1998), Smith (2000), Badici (2008) and Zhong (2012)) and dialetheists (principally Priest) have replied in turn (for example,

Priest (1998), Priest (2000), Weber (2010c) and Priest (2012)). Priest has also attempted to extend further the scope of the Inclosure Schema to encompass the paradoxes of vagueness (for example, Priest (2010b)) and to paradoxes arising from indefinitely extensible concepts (Priest 2013).

The debate concerning the Inclosure Schema is ongoing, and it is not my intention to adjudicate it here. What is clear, though, is that, if Priest is correct that it provides an accurate account of the paradoxes, and that like problems demand like solutions, then, firstly, his arguments for dialetheism about the semantic paradoxes (especially the revenge problem) extend far beyond the semantic paradoxes. Secondly, this puts pressure on competitor solutions to the paradoxes (such as Field's paracomplete account defended in (2008)) to develop similarly unified accounts of the paradoxes of inclosure.

3.4 Summary of Priest's Case for Dialetheism

This chapter has considered only some of the arguments presented by Priest for dialetheism. A number of other arguments have been made (by Priest and others) for the view. Many of these involve defending the usefulness of the applications of dialetheism to various problems in philosophy. If dialetheism can provide useful solutions to outstanding philosophical problems, this provides further, indirect support for the view. A number of these are discussed in *In Contradiction*: for example Priest argues that, in metaphysics, dialetheism offers an account of change (and, in particular, motion) which allows for genuine instants of change (see chapters 11 and 12 of *In Contradiction*, and chapter 15 for a related discussion of the metaphysics of time). In the philosophy of law, Priest thinks, dialetheism offers the best account of inconsistent legal systems, by treating the apparent contradictions arising therein as bona fide dialetheia.

The central arguments of Priest's other book-length treatments are also sometimes of this kind. His third book, for example, *Towards Non-Being* (2005) defends a view he calls 'noneism', a development of the view of Richard Routley (later Sylvan) given in his *Exploring Meinong's Jungle* (1980), itself a development of the views presented in various works by Alexius Meinong, especially his *The Theory of Objects* (1960). Noneism is the view that some

things don't exist and is put to work as a solution to a cluster of puzzles involving intentionality, as well as an account of mathematical objects, fictional objects and possible worlds. The biggest obstacle to this view, according to Priest (chapter 8 of his (2005) can be consulted for the details), is that a paradox of multiple denotation, to which the view appears subject, delivers the rather worrisome conclusion that 0=1. The argument for this unpalatable conclusion crucially requires the substitutivity of identicals, which fails in cases of multiple denotation, on the dialetheic semantics Priest provides. So a suitably dialetheic development of noneism, Priest hopes, can deliver an attractive solution of a number of long-standing philosophical problems concerning both intentionality and ontology.

Priest's most recent book *One* (2014) further expands dialetheism's potential applications in metaphysics, giving inconsistent accounts of unity, mereology, identity and instantiation, among other notions. These are interesting parts of the positive case for dialetheism, deserving of serious engagement and scrutiny which they don't receive here.

One reason for this is that it's fair to say that most philosophers regard the case for dialetheism based on the paradoxes, as discussed above, as the most powerful and persuasive case available for the view[3]. Another is that my concern in this thesis is with the problem of revenge. My aim in this chapter has been to sketch the case for dialetheism, based on the paradoxes, as it has been defended by Priest, and to emphasise the centrality of revenge to this case. The purpose of this is to illustrate the place of the thesis in the debate surrounding dialetheism, and to give a sense of the importance of my claims in this regard, if they are shown to be correct.

The case for dialetheism summarised in the other sections of this chapter is primarily negative, in the sense that it is concerned to rule-out alternatives to dialetheism by demonstrating their theoretical inadequacy. The first arguments discussed concerned the semantic paradoxes, paradigmatically, the Liar. The broad strategies considered in response to these paradoxes were Paracompletism

[3] It's possible that Priest, himself, is an exception to this, since he has said (*25 Years In Contradiction* conference, Glasgow, 2012) that the apparent contradictions arising from inconsistent laws have always struck him as the clearest cases of dialetheia. He may, of course, still think that the most *persuasive* case for his view is still made on the basis of the paradoxes.

(defenders of truth-value ‘gaps’), rejection of the T-scheme (which overlaps with the discussion of paracompletism) and hierarchical approaches (hierarchies of truth predicates, and hierarchies of interpretations, respectively). In the arguments against each of these, expressive limitations and, in particular, revenge problems, played a crucial role: the inability of paracompletism to express the notion of truth-value gap and the inability of hierarchical theories to express claims about the hierarchy as a whole are, in Priest’s view, devastating revenge problems.

The second class of arguments concerned the paradoxes of set-theory. Again, considerations of expressive limitation loom large. Priest attacks the underlying motivation for the view that the cumulative hierarchy exhausts the universe of sets in a number of way. But chief among the objections to this picture is that, in order to avoid the paradoxes, the view must disallow certain totalities, such as the set of all sets, which seem required for a proper interpretation of our practices. For example, we seem to be able, in Category Theory, to carrying out operations on the totality of the sets, but the only way to treat this behaviour in accordance with the picture of the sets given by the cumulative hierarchy is to interpret it as being about some initial segment of the cumulative hierarchy, rather than the universe of sets itself. In other words, we have seem to have expressive abilities, exemplified in the work of category theorists, which, on pain of triviality, defenders of the cumulative hierarchy must deny.

Finally I discussed Priest’s Inclosure Schema, which is an attempted characterisation of the underlying structure of the paradoxes. Though this is not directly related to revenge, in combination with the Principle of Uniform Solution, the Inclosure Schema promises to extend the reach of Priest’s arguments for dialetheism in the realm of the semantic paradoxes (especially, of course, revenge) much further afield. If he is right that the Inclosure Schema captures the root of the paradoxes, and the Principle of Uniform solution is correct, the problem of revenge may compel us to endorse dialetheism about a number of subjects, such as set theory, indefinitely extensible concepts, vagueness and perhaps others.

It should be clear now that as far as the case for dialetheism arising from the paradoxes goes, the problem of revenge is of singular importance. Dialetheism is

argued to be the only available view of these paradoxes because it is the only view to avoid the problem of revenge. So long as dialetheism can be thought to avoid the problem of revenge, this argument is clearly powerful. But if, as I claim, dialetheism too suffers from revenge, the argument loses its bite. Arguably, dialetheism is no *worse* off as regards revenge than any other view since, (again, arguably) they all seem subject to the problem too. But, if we agree that accepting true contradictions is theoretically more costly than, say, accepting violations of the law of excluded middle, or stratifications of the truth predicate, as I suspect most philosophers (though Priest may be an exception) would, then if dialetheism is to remain a tenable view this extra burden must be borne elsewhere. I suspect many philosophers in this area think that this burden could only be met by immunity from the revenge problem and, if this can't be had, dialetheism is untenable. This is not my view. Though I think it clear that a revenge problem is a blow to dialetheism (indeed, a serious one), I do not think dialetheism need be thought to stand or fall by revenge immunity. The set-theoretic arguments given by Priest, for example, still stand, and may be the best way for dialetheism to be motivated, if they do face revenge.

3.5 Priest's Dialetheism

This section describes Priest's version of dialetheism. I begin with his account of truth and falsity, then present the semantics underlying his theory and finally discuss the view's paraconsistent metatheory. As in 3.1, I follow Priest's presentation of things closely.

Priest begins his account of truth with a defence of the T-scheme (2006, p.55). As already discussed, one reason to accept the T-scheme is its important role in allowing generalisation. Another, according to Priest, is its importance for a theory of meaning. So long as one accepts Frege's observation that to give the meaning of a sentence is to give its truth conditions, truth is obviously central to a theory of meaning. As Priest says, one might, as Davidson (1967) does, think that a theory of meaning is just a Tarski-type construction of the kind discussed above and that the meaning of α is given by the relevant instance of the T-scheme, $T\langle\alpha\rangle \leftrightarrow \alpha$. Alternatively, he suggests (citing Montague (1974), Lewis (1970) and Routley *et al*. (1982)), one might give truth conditions not in terms of

truth *simpliciter*, but truth-in-a-possible-world, where α's meaning is given by the relevant instance: <*α*> is true in world *w* ↔ *α*, where *α* is parameterised to *w*. This is an instance of a more general scheme of which the T-scheme is the more specific case where *w* is taken to be the actual world. Even verificationists (See Dummett (1978)), who understand truth constructively, accept the T-scheme and that it gives sentences their truth-conditions and, thereby, their meaning. So, Priest thinks, the T-scheme must hold for any meaningful sentence, since it is a necessary part of the specification of its meaning.

Priest accepts that the T-scheme gives an extensional characterisation of truth, but thinks that it might provide more than this, depending on what we take its conditional to be. On the dialetheic semantic theory defended by Priest, the material conditional does not validate modus ponens. But the conditional in the T-scheme, Priest says, certainly seems to detach. So a dialetheist will need to add some other (intensional) conditional to their theory which supports detachment: what Priest calls a 'genuine' conditional. This gives us reason to hope, Priest thinks, that the T-scheme will provide more than a merely extensional account of truth.

However, Priest thinks, there must be more to truth than the T-scheme alone, even with the conditional understood intensionally. The argument for this, which he attributes to Dummett (1978, p. xxi), is as follows. According to Priest, to give the meaning of a sentence is to give its truth-conditions, which are given for a sentence *α* by the relevant instance of the T-scheme, *T*<*α*> ↔ *α*. This being the case, Priest claims, this instance of the T-scheme cannot be thought to provide *both* the sense of *α* and *α*'s truth-conditions. He explains this by supposing that someone has no grasp either of truth or of the senses of the sentences of some language *L*. If all such a person knows of *T*, and of the senses of the sentences of *L*, is that all the instances of the T-scheme hold, then they are not in a position to infer anything about either. For example, he says, the T-scheme would be satisfied if the sense of every sentence of *L* were its negation and *T* was taken to be a falsity-predicate. He illustrates with an example: consider a propositional language containing the connectives ∧, ∨ and ¬ and formulate its truth theory with every instance of *T*<*p*> ↔ *p'* as an axiom (where

p' is a translation into English of *p*, if *L* is not English, and simply *p*, if *L* is English) and the following recursive clauses for the connectives:

$T\langle \alpha \wedge \alpha \rangle \leftrightarrow T\langle \alpha \rangle$ and $T\langle \alpha \rangle$

$T\langle \alpha \vee \beta \rangle \leftrightarrow T\langle \alpha \rangle$ or $T\langle \beta \rangle$

$T\langle \neg \alpha \rangle \leftrightarrow$ it is not the case that $T\langle \alpha \rangle$

Though, for this language, we can prove every instance of the T-scheme, we cannot, says Priest, take this to have determined the senses of the sentences of the language, or that *T* is a truth-predicate; for the theory is compatible with *T* being a falsity predicate for the language, with every atomic sentence having its negation as its sense, with '$\wedge$' meaning 'or' and with '$\vee$' meaning 'and'. Since meaning and truth are mutually dependent, he thinks, we can fix the nature of one and find out about the other, but if neither is fixed, no information about either is available.

Hence, on Dummett's view, endorsed by Priest, the truth definitions specifying when an arbitrary sentence of the language is true cannot simultaneously provide an account of the meaning of each sentence in the language, unless we already know what the purpose of the predicate, *T*, is supposed to be. So the T-scheme on its own does not provide an adequate characterisation of truth.

There are a couple of points here with which one might take issue. Firstly, the Dummettian argument Priest gives does not directly establish that there is more to truth than the T-scheme. If it is correct, it establishes that one cannot *both* take there to be nothing more to truth than the T-scheme *and* take meaning to be nothing more than giving truth conditions. So though Priest takes the upshot of the argument to be that there must be more to truth than the T-scheme, one might equally, without independent arguments to the contrary, take it to show that there is more to meaning than truth conditions; perhaps what's required is a teleological account, not of truth, but of meaning.

Secondly, the move from these considerations directly to the claim that what's required for a grasp of meaning is an understanding of the *purpose* of the truth predicate is a *non sequitur*. At best what has been established is that *something*

extra is required in our characterisation of truth. Whether this is a characterisation of the purpose of the predicate (as in Priest's teleological account described below) is a further claim that must be argued for.

So, at the least, it seems that further argumentation is required here. But supposing this can be given, we now need an account of truth which goes beyond that offered by the T-scheme and which characterises the purpose of the truth predicate. Priest's preferred account is what he calls the 'Teleological Account' of truth, and I turn to this now.

3.5.1 The Teleological Account of Truth

Priest's view is that the purpose (*telos*) of assertion is to say something true and that this is the essential fact about truth missed by the T-scheme (2006, p.61). He attributes this view to Dummett ((1973), p.320) and also approvingly endorses Dummett's analogy comparing assertion to the playing of a game, and of speaking the truth with winning. Merely knowing what counts as a winning position in, say, chess, without knowing that this position is what players of chess are aiming at, is insufficient for knowing what winning is. Similarly, according to Priest, merely knowing what it is that makes each sentence true is insufficient for grasping the nature of truth. We must also know that truth is the aim of assertion.

He admits that this is not to say that every speaker aims all the time at speaking the truth, just as it may not always be the aim of every player of every game to win. Someone may assert with the intention of deceiving, or play a game with the intention of losing. Nonetheless, he thinks, this does not change the fact that the point of asserting and of playing games as such are, respectively, speaking the truth and winning.

One reason he gives for thinking this is the way in which abstract theories of truth are tested (citing discussions by Davidsonians, and suggesting that similar tests would apply to Montague grammar). According to Priest, this is done (given that the meaning of α is supposed given by the instance of the T-scheme $T\langle\alpha\rangle \leftrightarrow \alpha$) by seeing whether speakers are disposed to assert α "…when they may

reasonably be taken to believe that α' (or at least may reasonably be taken to intend the hearer to believe (that they believe that) α')."(2006, p.62).

This, according to Priest establishes, or at least supports, the claim that "…it is the use to which the truth predicate is put, and in particular its connection with the things that speakers wish to or are prepared to assert, that completes its characterisation." (2006, p.62).

It's not clear that the fact of the use of these techniques really helps Priest's case. Firstly, is not normally taken to straightforwardly follow from the fact that something forms part of our method of finding out about a particular thing that it is thereby a part of the correct characterisation of that thing. No one but the most extreme instrumentalist thinks, for example, that an account of telescopes is an essential part of the proper characterisation of what it is to be a star.

Secondly, what those testing theories of truth are aiming at is discovering whether natural language is such that the meaning of α is given by $T\langle\alpha\rangle \leftrightarrow \alpha$. The most relevant piece of information, in this regard, is whether speakers' actual linguistic practices, and the contents of their beliefs (and perhaps other cognitive states), are in accord with this; that is, whether speakers take α to be true if and only if they believe α. The best way of finding out whether this is the case is to look at speakers' assertions as a guide to when they think 'It's true that α and, obviously, to consider only those cases where we may reasonably believe that the phonemes uttered are such that the 'α' within the scope of the truth-predicate is actually supposed to pick out α.

Whether this is a reasonable way of testing the veracity of truth-conditional semantics is independent of whether the teleological account of truth is correct, and so the fact that this is the methodology used by Davidsonians to check theories of meaning provides no support to the teleological account of truth. This is partly evidenced by the fact that neither Davidson nor Davidsonians are themselves proponents of a teleological theory of truth.

Another way one might object to Priest's account (see Weir (2004)) is to deny that assertion has any *telos* at all: like any linguistic act, assertion is put to a

wide range of uses (to deceive, to amuse, to make money etc.) and there is no good reason to prefer one as fundamental.

3.5.1.1 Falsity

The most important part of Priest's account of truth, from the point of view of revenge, is that he uses it to motivate an asymmetric treatment of truth and falsity and, in turn, to reject the principle that falsity entails untruth. This makes a significant difference to Priest's position with respect to various revenge objections often thought to beset dialetheism, so I will now give an account of Priest's views on the matter.

As is standard, Priest defines falsity in terms of truth and negation via the following schema:

$F\langle\alpha\rangle \leftrightarrow T\langle\neg\alpha\rangle$

He notes (2006, p.64) that with the obvious definition of negation as a sentential function taking a true sentence to a false one and vice versa, combined with the above definition of falsity, results in circularity. But he also points out that this problem is not unique to dialetheism: in general, negation and falsity are definable only in terms of one another. There are, however, general principles concerning the notions, on which Priest takes a view. Firstly, he accepts what he calls 'exhaustion', that untruth entails falsity:

$\neg T\langle\alpha\rangle \rightarrow F\langle\alpha\rangle$

He sees this principle as following from the teleological account of truth. In a two-player game, a draw is possible; but in a one-player game, one either succeeds in winning, or one fails. Assertion, according to Priest, is a one-player game. The goal of asserting is to speak truly and anything less than this is failure: there is "no third possibility" (2006, p.64).

I find the principle that untruth entails falsity extremely compelling. Nonetheless, it's unclear that Priest's discussion here really motivates it. Firstly, when he talks of a "third possibility" here, what he has in mind is truth-value gaps. But if the existence of truth-value gaps are thought to motivate rejecting

the principle, then it is this, not the teleological account of truth, which is doing the work. The gappist could, after all, perfectly well accept the point that anything less than truth counts as 'losing' at assertion, but think that there are two ways to lose: by asserting a falsehood, or by asserting a gap. A similar point is made by Parsons (1990, p.342).

But, in fact, it is not clear that admitting truth-value gaps need violate the principle at all. If we take gap theory to be committed to the claim that some sentences are neither true nor false, then the principle can be easily seen to fail. But there is an important reason for gappists to avoid putting their view in this way: it takes very little to demonstrate that it leads to inconsistency. But if it is not a consequence of a gappist theory that gap sentences are untrue, then there is no reason that I can see to reject the principle that the sentences which *do* receive the value untrue should also be false. In any case, to reject the principle is a further claim and one that a gappist need not endorse. On the other hand, is hardly a virtue of gappist theories that they cannot assert that sentences receiving the gap value fail to be true (since those sentences, plausibly, do fail to be true).

In a brief discussion of the claims, related to my own, made by Parsons, Priest concedes the point that it his acceptance of the law of excluded middle plays an important role in the argument (2006, pp.267-268). But, he says, the argument is still significant, because it shows that there is nothing flowing just from considerations of assertion which distinguishes between the ways in which an assertion may fail: and so nothing "…for the distinction between α being false and α being neither true nor false to get a grip on." (2006, p.267) It therefore raises a problem, he thinks, for gappist theories: they must find something which grounds the distinction between being false and being neither true nor false. He expresses scepticism that this can be done.

Setting aside the concerns already mentioned about this way of putting the gappist view, this is a problematic point for Priest to be making. He goes on, as is important for his view, to defend the claim that there is an important difference between being false and being untrue. But if there is nothing in assertion to distinguish being false from being neither true nor false, then equally, there seems nothing in assertion to distinguish being false from being

untrue. Each seems to count merely as a failure, as far as Priest's view goes, and the prospects for grounding the distinction elsewhere seem as good (or bad) as the gappist's for grounding their distinction. I return briefly to this issue in the next section.

3.5.1.2 Untruth

Another principle one might expect to hold, which Priest calls 'exclusion', relates falsity with untruth:

$F\langle\alpha\rangle \rightarrow \neg T\langle\alpha\rangle$

Priest rejects this principle (2006, pp. 70-72). One reason, he suggests, to think the principle holds is that it follows from the contraposed T-scheme combined with the transitivity of the conditional. This argument fails for Priest, since the conditional he takes to occur in the T-scheme does not contrapose.

One reason he considers that it does not hold, at least for dialetheists, is that there are some things, viz. dialetheia, which are false but are also true, and so it we ought not to think falsity entails untruth. This argument, too, is unpersuasive, at least for Priest's version of dialetheism since, as he says, if we have a sentence which is both true and false, exclusion delivers the result that the sentence is both true and untrue. But this sort of contradiction is perfectly acceptable by the lights of Priest's dialetheism. There are also strengthened liar sentences, predicating untruth of themselves, which Priest accepts (independently of exclusion) are both true and untrue. So exclusion is not, as might be thought, incompatible with dialetheism, at least of the kind endorsed by Priest.

Though contradictions of this sort are perfectly acceptable for Priest, one effect of accepting exclusion would be to turn every 'internal' contradiction (one which is true and false) into an external one (which is true and untrue). The main reason Priest gives for rejecting exclusion is that this feature of it multiplies contradictions beyond necessity which, he argues (2006, p.115), we should not do.

A thorough discussion of this methodological maxim is beyond the scope of this thesis, so I restrict myself to a few brief remarks. Firstly, accepting exclusion does not multiply contradictions in the sense of generating a new, external, kind of dialetheia, previously absent. As I have said, Priest already accepts that the strengthened liar is both true and untrue, independently of exclusion.

Moreover, once we have one such sentence, we can generate infinitely many. So neither does exclusion multiply contradictions in the sense of increasing the cardinality of the set of external contradictions.

One reason Priest gives for the principle is that dialetheia have a low statistical frequency (2006, p.116) and that this is, generally, though of course defeasibly, grounds for rejecting them (2006, p.102). But the statistical frequency of dialetheia will vary depending on the discourse (Priest asks us to do a 'head-count' of randomly chosen assertions recently encountered, and suggests that if it's high, we've been reading about dialetheism (2006, p.116)). But the effect of exclusion is to turn internal contradictions into external ones, and so the discourse we're considering is, perhaps, semantics, or the metaphysics of change, and already granted to be rife with inconsistency. So, it seems to me, we have reason to be at least cautious of the maxim and the claim that it tells against exclusion.

In any case, as above, the principle that contradictions not be multiplied is defeasible, and considerations in favour of exclusion may well be taken to outweigh the maxim, even if we take it to carry weight in this case. I mention two potential reasons in its favour.

The first is that, without exclusion, the behaviour of untruth is seems problematically underdetermined. From the exhaustion principle, we know that anything untrue is thereby false, but we don't have principles determining when a falsity is, additionally, untrue. For example, a strengthened liar sentence, *L*, predicating untruth of itself is accepted by Priest to be untrue. One would therefore expect, for example, the double-negation of *L* to also be untrue. But there is no way that I can see of showing this to be the case, on Priest's view. Indeed, without any principle like exclusion, it is difficult to see what, in Priest's theory, is determining the matter.

Since falsity is defined as truth of negation, the same point applies to the less-discussed notion of lacking falsity which we might call 'unfalsity'. For any sentence α which is unfalse, $\neg T\langle\neg\alpha\rangle$, I see no way in general of obtaining the result that its double-negation, $\neg T\langle\neg\neg\neg\alpha\rangle$ is unfalse, and again, see nothing which determines whether this is so. These issues arise again in the context of my own revenge problem for Priest's view.

The second reason is, as mentioned above, that, if Priest's account of truth is as the telos of assertion, and he wishes to defend that falsity and untruth are distinct, then it would be desirable for this asymmetry to be grounded in our practices of assertion (at least as he understands those practices). But, from the point-of-view of Priest's characterisation of assertion, there is no difference between falsity and untruth, since both are just equally unsuccessful ways of failing to achieve the goal of speaking the truth. As I have said, Priest himself takes a sceptical view on whether such a distinction, in the case of gappists, can be grounded elsewhere.

Priest accepts that the differences between falsity and untruth are "surprisingly little" (2006, p.71) and provides, as the only example of their distinctness, the following. The strengthened liar is both true and untrue. It follows that:

$$\exists x(T\langle x\rangle \wedge \neg T\langle x\rangle)$$

But by the law of excluded middle (or at least an instance of it), $\forall x(T\langle x\rangle \vee \neg T\langle x\rangle)$, the De Morgan equivalencies and the quantifier rules, it follows that:

$$\neg\exists x(T\langle x\rangle \wedge \neg T\langle x\rangle)$$

So there both are and are not truths which are untrue. In the case of falsity, it is certainly true (dialetheically) that $\exists x(T\langle x\rangle \wedge F\langle x\rangle)$, but it is not (or at least not obviously) the case that $\neg\exists x(T\langle x\rangle \wedge F\langle x\rangle)$. So, Priest thinks, untruth is 'more inconsistent' than falsity. This seems true enough, as things stand, but it is clearly not the sort of difference which would ground a distinction between falsity and untruth sufficient to motivate rejecting exclusion. This is because the only reason that this difference between falsity and untruth exists is that

exclusion has already been taken to fail: if we add exclusion, falsity and untruth are logically equivalent and $\neg\exists x(T\langle x\rangle \wedge F\langle x\rangle)$ follows.

3.5.2 Formal Semantics for Priest's Dialetheism

This section describes the formal semantics underlying Priest's theory, beginning with the connectives Priest takes to be extensional, followed by the conditional, which he takes to be intensional. The presentation, in *In Contradiction*, of the semantics for Priest's preferred logic, *LP*, treats an interpretation, as is standard, as a function from a formula to a set of truth values. However, Priest has revised his view and now treats an interpretation as a relation between a formula and a truth value. So in this section, I follow his presentation of the semantics given in his (2001, Section 8.2).

3.5.2.1 Extensional Connectives

Priest describes his semantics as follows (2001, pp.142-144): we consider, initially, a propositional language with the set of parameters, *P*, whose set of formulae, *F*, is the closure of *P* under $\wedge$, $\vee$ and $\neg$. We can define a material conditional, $\alpha \supset \beta$, by $\neg\alpha \vee \beta$. An evaluation is a relation, ρ, between a formula a truth value, $\rho \subseteq P \times \{1,0\}$. With no restriction on ρ other than that every member of *P* be related at least one member of π, the recursive clauses for the connectives are as follows:

(1a) $\neg\alpha\rho 1$ iff $\alpha\rho 0$

(1b) $\neg\alpha\rho 0$ iff $\alpha\rho 1$

(2a) $(\alpha \wedge \beta)\rho 1$ iff $\alpha\rho 1$ and $\beta\rho 1$

(2b) $(\alpha \wedge \beta)\rho 0$ iff $\alpha\rho 0$ or $\beta\rho 0$

(3a) $(\alpha \vee \beta)\rho 1$ iff $\alpha\rho 1$ or $\beta\rho 1$

(3b) $(\alpha \vee \beta)\rho 0$ iff $\alpha\rho 0$ and $\beta\rho 0$

As Priest points out, these are essentially the conditions of classical semantics, except that in that case, the exclusivity of truth and falsity make the second condition of each pair redundant.

We then define logical consequence as follows: If $\Sigma \subseteq F$ and $\alpha \in F$, then

$\Sigma \vDash \alpha$ iff for every interpretation, ρ, if $\beta\rho 1$, for all $\beta \in \Sigma$, then $\alpha\rho 1$

And logical truth as:

$\vDash \alpha$ iff for every interpretation, ρ, $\alpha\rho 1$

Some important features of *LP* (refer to (2006, pp. 80-81) for proofs) are:

1) $\vDash \alpha$ iff α is a logical truth of classical logic
2) If $\Sigma \vDash \alpha$ then α is a classical consequence of Σ
3) It is not the case that, if α is a classical consequence of Σ, $\Sigma \vDash \alpha$. Importantly, $\alpha \wedge \neg\alpha \vDash \beta$ fails, as does $\neg\alpha$, $\alpha \vee \beta \vDash \beta$

The third fact is of particular importance for dialetheism, since the invalidity of these principles is required for avoidance of triviality, and is what makes the logic 'paraconsistent'.

We add quantifiers to the picture in the following way (2001, pp.467-469). An interpretation is a pair $<D, v>$, where D is the (non-empty) domain of quantification. For every constant term in the language c, $v(c) \in D$, for every n-place predicate, P, $v(P)$ is the pair $<E,A>$, where E and A are subsets of D^n and understood to be the extension and anti-extension of P, respectively. We can write these as $v^E(P)$ and $v^A(P)$, respectively.

We can then introduce the relation, ρ, between formulae and truth-values recursively with the following conditions:

$P\alpha_1 \ldots \alpha_n \rho 1$ iff $<v(\alpha_1),\ldots,v(\alpha_n)> \in v^E(P)$

$P\alpha_1 \ldots \alpha_n \rho 0$ iff $<v(\alpha_1),\ldots,v(\alpha_n)> \in v^A(P)$

The conditions for the connectives are the same as the propositional case above. The conditions for the quantifiers are:

$\forall x A\rho$ 1 iff for all $d \in D$, $A_x(k_d)\rho 1$

$\forall x A\rho 0$ iff for some $d \in D$, $A_x(k_d)\rho 0$

$\exists x A\ \rho 1$ iff for some $d \in D$, $A_x(k_d)\rho 1$

$\exists x A\rho 0$ iff for all $d \in D$, $A_x(k_d)\rho 0$

Negation interacts with the quantifiers as follows:

$\neg\forall x A\rho 1$ iff $\exists x \neg A\rho 1$

$\neg\exists x A\rho 1$ iff $\forall x \neg A\rho 1$

As in the propositional case, validity is defined as preservation of value 1:

$\Sigma \vDash A$ iff in every interpretation where $B\rho 1$ for all $B \in \Sigma$, $A\rho 1$

The following two constraints are available:

Exclusion: for every *m*-place predicate, *P*, and $d_1,\ldots,d_m \in D$, $\langle d_1,\ldots,d_m\rangle \notin v^E(P) \cap v^A(P)$

Exhaustion: for every *m*-place predicate, *P*, and $d_1,\ldots,d_m \in D$, $\langle d_1,\ldots,d_m\rangle \in v^E(P) \cup v^A(P)$

If the first is imposed, but not the second, we obtain quantified K_3. If we impose the Exhaustion but not Exclusion, we obtain quantified *LP*, as desired. If we impose both, classical logic results.

3.5.2.2 Priest's Conditional

The logic just specified is *LP*. As already mentioned, we can define a material conditional for *LP*, '$\supset$', such that 'A $\supset$ B' is equivalent to '$\neg$A $\vee$ B'. But with the conditional defined in this way, modus ponens, A $\supset$ B $\vDash$ B, is equivalent to disjunctive syllogism, A $\vee$ B, $\neg$A $\vDash$ B, and disjunctive syllogism is invalid in *LP*, and so modus ponens is invalid in *LP* for a material conditional. There are two options here, for the dialetheist endorsing *LP*: they can either accept that *modus ponens* fails for the theory's conditional, or they can add a conditional to the language which does satisfy modus ponens.

The first option, to accept the invalidity of *modus ponens*, has not proved attractive with dialetheists (though see Beall (2013)), for the obvious reason that *modus ponens* is not only essential to our reasoning, both theoretical and everyday, but seems an essential part of what a conditional *is*. It does not seem unreasonable to insist that, if our definition of '→' is such that *modus ponens* fails, then '→' is not a conditional. Priest himself (2006, p.83) thinks it analytic of implication that it satisfies the principle.

So Priest chooses the second option and introduces a conditional satisfying *modus ponens* into the language. Important constraints are placed on this account by Curry's paradox. One way, which Priest favours, is to reject the principle of *contraction*:

$\alpha \rightarrow (\alpha \rightarrow \beta) \vDash \alpha \rightarrow \beta$

Which he puts in the form of the principle he calls 'assertion':

$((\alpha \wedge (\alpha \rightarrow \beta)) \rightarrow \beta$

Priest characterises the paradox as arising from a sentence, C, equivalent to:

$T\langle C\rangle \rightarrow \beta$

The T-scheme instance for C gives us:

(1) $T\langle C\rangle \leftrightarrow (T\langle C\rangle \rightarrow \beta)$

Supposing assertion holds, then, by the substitutivity of equivalents:

(2) $(T\langle C\rangle \wedge T\langle C\rangle) \rightarrow \beta$

And hence, by the properties of conjunction:

(3) $T\langle C\rangle \rightarrow \beta$

By (1) and modus ponens, T<C>, and, by (3) and modus ponens, β. Hence, according to Priest, any logic containing assertion is trivial. This is true, of course, only if the view in question also contains the other things involved in the proof of triviality, but certainly Priest's view contains these principles, and so, for him (and many others), assertion will have to be avoided.

The conditional favoured by Priest is a strict, non-contraposable conditional. He augments the semantics already provided in the following way (2006, p.270-273). An interpretation contains two disjoint sets of worlds, *P* and *I*, which are the logically possible worlds and the logically impossible worlds (the worlds at which the laws of logic differ from the actual), respectively. Validity is defined as truth preservation at the possible worlds.

The conditions for the extensional connectives carry-over unchanged except that they are relativized to worlds. For example, the conditions for conjunction are now:

$(\alpha \wedge \beta)\rho_w 1$ iff $\alpha\rho_w 1$ and $\beta\rho_w 1$

$(\alpha \wedge \beta)\rho_w 0$ iff $\alpha\rho_w 0$ or $\beta\rho_w 0$

Since the law of excluded middle is a law of logic, we add as a constraint on the logically possible worlds, *P*, (but not on the impossible worlds, whose laws may differ) that:

(*) For all $w \in P$, $\alpha\rho_w 1$ or $\alpha\rho_w 0$

The truth and falsity conditions for the conditional are now, where $w \in P$:

$(\alpha \rightarrow \beta)\rho w 1$ iff for all $x \in W$, if $\alpha\rho_x 1$, then $\beta\rho_x 1$

$(\alpha \rightarrow \beta)\rho w 0$ iff for some $x \in W$, if $\alpha\rho_x 1$, then $\beta\rho_x 0$

These conditions make $p \rightarrow p$ a logical truth, but at impossible worlds, logical truths may be different. Priest's preferred way to represent this (2006, p.271) is to invoke the ternary relation, *R*, on worlds of Routley/Meyer semantics. If $w \in I$, the truth/falsity conditions of the conditional are then:

$(\alpha \rightarrow \beta)\rho_w 1$ iff for all $x,y \in W$, such that $Rwxy$, if $\alpha\rho_x 1$, then $\beta\rho_y 1$

$(\alpha \rightarrow \beta)\rho_w 0$ iff for some $x,y \in W$, such that $Rwxy$, $\alpha\rho_x 1$ and $\beta\rho_y 0$

In fact, Priest takes these conditions to give the truth/falsity conditions for '→' uniformly, but with the restriction on *R* that, if $w \in P$, then $Rwxy$ holds iff $x = y$, we obtain the simplified conditions given previously (2006, p.271n25). These

conditions now explain how $p \rightarrow p$ can fail at impossible worlds: we consider a case where *Rwxy* and p holds at x, but not at y.

We can also, Priest says, see how contraction fails: consider some $w \in P$, such that $p \rightarrow (p \rightarrow q)$ holds at w, and *Rwxy*. Suppose x is some world such that p holds there and, hence, $p \rightarrow q$ holds there too. It does not follow that q also holds at x, since x might be an impossible world at which *modus ponens* fails. We can also see that the conditional fails to contrapose: consider some world w such that $p \rightarrow q$ holds at w, and *Rwxy*, and p is true only at x, but q is both true and false at x, so $\neg q \rightarrow \neg p$ fails to hold at w. This follows from the fact that it is not stipulated of the conditional that it preserve falsity backwards.

Thus far, nothing in the above conditions ensures that excluded middle holds, though it is Priest's view that it does. His preferred way of avoiding this problem (2006, p.272) is to add a principle he calls the 'Augmentation Constraint', according to which:

AC: if α is neither true nor false at w, there are worlds, w_0 and w_1, such that $[w] \subseteq [w_0]$ and $[w] \subseteq [w_1]$, α is false at w_0, and true at w_1

Where $[w]$ is understood as the set of all those things true at w. The principle, according to Priest, captures the thought that, for any world containing a truth-value gap, that gap could be filled, either with a falsity or with a truth. So, for any world containing a gap, there exist worlds, essentially the same, but with the gap filled with a truth/falsity. For example, if there were some world w, at which $p \rightarrow q$ were neither true nor false, then there would be some other world at which p is true (only), but q is neither true nor false. But then, by *AC*, there will be some further world at which p is true (only), but q is false and, hence, $p \rightarrow q$ is false (not neither true nor false) at w after all.

Priest notes that whether *AC* holds given the specification of an arbitrary interpretation is not, in general, effectively checkable. Though, "[w]hether there are conditions on the components of an interpretation that are necessary and sufficient to realise the constraint, and which can be effectively checked, is not, at the time of writing, known" (2006, p.272). One thing to which we can appeal, however, is the existence of the trivial world, $w_\perp$. At $w_\perp$, everything is

both true and false, and the only instance of R into which $w_\perp$ enters is $Rw_\perp w_\perp w_\perp$. For any world, w, and any propositional parameter, a, $w_\perp$ plays the role of both w_0 and w_1. On the effects of the trivial world, $w_\perp$, Priest says, "The presence of $w_\perp$ has a very strong effect on conditionals at possible worlds: they are all at least false; but the behaviour of negated conditionals is often not very important." (2006, p.273)

So, for any possible world, w, and any propositional parameters p and q, p is true at $w_\perp$ and q is false at $w_\perp$, and so $p \rightarrow q$ is false at w. So, in fact, every formula containing a conditional, at every possible world, including the actual, is false. All dialetheists accept that at least some things are both true and false, and almost all reject trivialism (though, for the one exception of which I am aware, see Kabay (2010)). Between these two, there is room for reasonable disagreement about how much inconsistency is too much. But it seems to me the conclusion that every conditional expression (at least, every expression containing the conditional above) is false is a very radical one indeed, and that it deserves serious discussion, and some explanation. It certainly doesn't seem that every true conditional is a true contradiction, nor does it seem that this being so follows obviously from dialetheism, so that its denial might be thought to arise from a question-begging refusal to take dialetheism seriously. The claim that the behaviour of negated conditionals is often not important is, even if granted, insufficient, since the behaviour of negated conditionals often *is* quite important too.

One problem for Priest which arises from this is that it puts pressure on his account of classical recapture. It's clear that some instances of, for example, disjunctive syllogism, are perfectly legitimate. If I know that someone is either in the park or the pub and discover that they are not in the pub, I am entitled to conclude that they are in the park. Priest agrees and proposes that the way for a dialetheist to account for this is to defend the legitimacy of reasoning by the (strictly invalid) disjunctive syllogism in consistent contexts (this is the goal of (2006), Chapter 8). Part of this account rests on the improbability of dialetheia, which "…appear to occur in a quite limited number of domains: certain logico-mathematical contexts, certain legal and dialectical contexts (which I will discuss in Part Three), and maybe a few others." (2006, p. 116)

But if every conditional is false and, hence, every true conditional is a dialetheia, this is no longer true. Dialetheia appear in *every* context in which conditionals occur and so occur with very great frequency in almost every discourse. This might be thought unfair, since it is not Priest's view that the logical conditional specified above is the conditional 'if…then…' in English (2006, p.273n29). Perhaps, then, the logical conditional and its inconsistency occur less frequently than might be thought, though some case ought to be made to this effect.

Priest's preferred account of the conditional in English is as a relevant *ceteris paribus* conditional (as specified in his ((2001), pp.84-85). Very roughly, on this view, $A \rightarrow B$ is true (at a world) just if, at every (accessible) world at which $A \wedge C_A$ holds (where C_A is an open-ended set of *ceteris paribus* conditions), B also holds. Whatever the *ceteris paribus* conditions require will trivially be satisfied at the trivial world and so, for any world w at which the trivial world is accessible, $A \wedge C_A$ will hold, but B will fail to hold, at the trivial world, and so $A \rightarrow B$ will be false at w. Whether the trivial world is accessible from the actual may depend, for example, on the restrictions placed on the accessibility relation: it will be accessible if, for example, the relation is reflexive, symmetric and transitive; but it may not be if the relation is weaker, though this has other costs. In any case, the trivial model at least *threatens* to falsify every such conditional at the actual world. This being the case, some account of which conditionals are falsified by the trivial model, why we should think this is acceptable and the extent to which this affects, for example, classical recapture, seems important.

Some facts (the first of which has already been noted above) which hold of Priest's conditional[4] (2006, p.86-87) are:

1. $\vDash \alpha \rightarrow \alpha$
2. $\vDash \alpha \leftrightarrow \neg\neg\alpha$

[4] In Priest's original (2006, pp.86-87) presents these in two distinct lists. The first is originally given as a list of principles holding of an entailment connective, of which another condition, contraposability, also holds. Since this doesn't hold of Priest's strict conditional, but the rest do, they are presented without the contraposability condition. On the second list, the condition $\{\alpha \leftrightarrow \beta\} \vDash \delta \leftrightarrow \delta(\alpha/\beta)$, where $\delta(\alpha/\beta)$ is δ with any subformula α replaced by β, appears (again concerning entailment), though I omit it since it may fail for the strict conditional in cases where δ contains a negation.

3. $\vDash (\alpha \wedge \beta) \rightarrow \alpha$
4. $\vDash \alpha \rightarrow (\alpha \vee \beta)$
5. $\vDash (\alpha \wedge (\beta \vee \gamma))$ iff $((\alpha \wedge \beta) \vee (\alpha \wedge \gamma)$
6. $\vDash ((\alpha \wedge \beta) \wedge (\alpha \wedge \gamma)) \rightarrow (\alpha \rightarrow \gamma)$
7. $\vDash ((\alpha \rightarrow \beta) \wedge (\alpha \rightarrow \gamma)) \rightarrow (\alpha \rightarrow (\beta \wedge \gamma))$
8. $\vDash ((\alpha \rightarrow \gamma) \wedge (\beta \rightarrow \gamma)) \rightarrow ((\alpha \vee \beta) \rightarrow \gamma)$
9. $\{\alpha, \beta\} \vDash \alpha \wedge \beta$
10. $\{\alpha, \alpha \rightarrow \beta\} \vDash \beta$
11. $\{\alpha \wedge \neg\beta\} \vDash \neg(\alpha \rightarrow \beta)$
12. $\{\alpha \rightarrow \beta\} \vDash (\gamma \rightarrow \alpha) \rightarrow (\gamma \rightarrow \beta)$
13. $\{\alpha \rightarrow \beta\} \vDash (\beta \rightarrow \gamma) \rightarrow (\alpha \rightarrow \gamma)$
14. $\{\alpha \rightarrow \neg\alpha\} \vDash \neg\alpha$

Now that the semantics for the conditional are fully specified, a slightly modified (constant domain) semantics for quantification and the identity predicate are added. The details can be found in ((2006) pp.92-93 and p.273). Now that the account of the semantics of Priest's theory has been specified, I turn to Priest's paraconsistent metatheory.

3.5.3 Paraconsistent Set Theory

As has already been said, perhaps the most important feature of Priest's theory, from the point-of-view concerned with revenge, is that it is taken to be characterised in dialetheic terms. That is, Priest's inconsistent theory is characterised in a paraconsistent (indeed, inconsistent) metatheory. This allows Priest to dispose of the object-theory-metatheory distinction as a significant distinction, and for the object language to describe its own semantics.

The effect of this, apart from the attractive consequence that the object theory (mirroring natural language) is able to describe its own semantics, is to block the argument for what I have called 'formal revenge'. The revenge problem is, in essence, the problem that a theory subject to the problem is unable to express some notion, expressible in natural language, due to certain extended liar-type paradoxes. The theory is therefore inadequate as a characterisation of how

natural language is not trivialised by the paradoxes, since natural language possesses the expressive powers the theory in question lacks. The formal version of this problem arises when we construct a notion in the account's metatheory which, were it expressible in the object theory, would result in the account's triviality. Since natural language contains this metatheory, the notion is *a fortiori*, expressible in natural language, though not in the account's object language, and we have a revenge problem.

Since, on Priest's account, the object language describes gives its own metatheory, anything expressible in the metatheory is, *a fortiori*, expressible in the object theory. The attempt to get formal revenge on Priest's theory, then, is just, in effect, the attempt to find a triviality proof for his theory. The prospects for this do not look terribly promising.

This, as we shall see, still leaves open the possibility of developing what I have called an 'informal revenge' problem for Priest's dialetheism. That is, we find some notion dialetheists use in their informal characterisation their view (and which they are thereby committed to the legitimacy of) and demonstrate that, due to certain revenge paradoxes, it cannot be expressed in the dialetheist's object theory.

If Priest's metatheory is to be paraconsistent, then it must be characterised in paraconsistent set theory. Priest discusses three possible strategies for developing paraconsistent set theory: The Material Strategy, The Relevant Strategy and The Model-Theoretic Strategy. Again, this discussion follows Priests (2006) closely, particularly Chapter 18.

As might be guessed from Priest's views on the set-theoretic paradoxes, discussed in, one thing common to each of the approaches he discusses to paraconsistent set theory is that they each contain:

(*Abs*) $\exists y \forall x\ (x \in y \leftrightarrow \alpha)$

(*Ext*) $\forall x(x \in z \leftrightarrow x \in y) \rightarrow z = y$

Priest endorses the further restriction that x not occur free in α (2006, p.248), though notes that this is optional (citing Routley (1980, p. 924 f.) as showing that the system proves choice and Brady (1989) as demonstrating its non-

triviality). As Priest says, paraconsistent set theory must find a balance between being too strong (proving triviality) and being too weak (so as not to be able to capture at least some reasonable amount of standard set theory). One important issue here concerns the conditional. If the conditional satisfies both *modus ponens* and contraction, then Curry's paradox delivers triviality (along with some other assumptions, like the standard structural rules). So the details of the conditional employed in the theory will be of particular importance.

One option is to employ a material conditional. On Priest's view (2006, p.249), this would involve employing the logic *LP* characterised earlier, but without the addition of a genuine conditional. Instead, on this strategy, the conditional is material, with $\alpha \supset \beta$ defined as $\neg\alpha \vee \beta$. Since, in *LP*, *modus ponens* fails for the material conditional (because *modus ponens* is equivalent to disjunctive syllogism, which must fail), this avoids Curry's paradox.

Priest points out (citing Restall (1992) for the details) that this system is non-trivial, validates all the axioms of *ZF* which are instances of (*Abs*), the Axiom of Infinity, but not the Axiom of Foundation and proves the existence of a universal set. What else is proved by the system (especially interesting theorems established in *ZF*) is, according to Priest, unknown, but the prospects do not look good. The importance of *modus ponens* to mathematical reason is enormous and its invalidity, on this strategy, is "a singular handicap" (2006, p.249).

He is slightly more hopeful that more can be established in the non-monotonic extension of *LP*, *LPm* (described in (2006, Chapter 16)). According to Priest (2006, p.251), it was not known at the time of his writing how much can be established in this way, and as far as I am aware, the situation has not changed. Since the approach still lacks a detachable conditional, we should be pessimistic.

According to Priest, another, more plausible option, is to carry out paraconsistent set theory in a logic containing a conditional which obeys *modus ponens*, but for which contraction fails. In particular, he suggests a 'depth relevant' logic, for which he provides a tableau system (2006, pp.251-255), and discusses the prospects for developing set theory in this logic.

Since the publication of these remarks in the second edition of *In Contradiction*, significant progress had been made on this strategy. In particular, Zach Weber has developed paraconsistent set theory using a relevant logic, and has proved a number of interesting results, including the Axiom of Choice, Cantor's Theorem, the existence of large cardinals, the falsity of the Continuum Hypothesis as well as the main results of ordinal and cardinal arithmetic ((2010a), (2010b) and (2012)).

The logic employed by Weber is not the same as that employed by Priest (one difference is that Weber's conditional is contraposable ((Weber) 2013, p.316)), but this still makes the relevant strategy a very promising option for a dialetheist development of set theory, though work on it is still in its early stages.

A further option returns to the theory of the material strategy, but rather than attempting to prove results directly in the theory, using the (non-detachable) material conditional, we construct models of the theory, which we call *M*. There are a number of models of the theory, many of which, Priest points out, are pathological e.g. the model containing a single set which both is and is not a member of itself.

More interesting models of the theory can be constructed using the Collapsing Lemma (details of exactly how these are constructed can be found in (2006, p.227 and pp.256-258)). Priest's preferred model is $\mathcal{M}^{-}$, which contains an initial segment of the cumulative hierarchy, V_{ξ}, up to some inaccessible ξ, as well as a single inconsistent set [*a*]. The interesting features of $\mathcal{M}^{-}$ from Priest's point of view are, firstly, that it is a model of naïve set theory (in particular, *M*), secondly, that it is, by the Collapsing Lemma, a model of *ZF* and, finally, that with quantifiers restricted to V_{ξ}, any theorem of *ZF* holds in $\mathcal{M}^{-}$ and, so, it contains an initial segment of *ZF* as a consistent inner model.

Priest says "We may therefore suppose that the true interpretation of the language of set theory has these properties…And to return, at last, to the question of what to make of the theorems of orthodox set theory, ZF, on this approach. The answer is obvious. Since the universe of sets is a model of ZF (as well as naive set theory), these hold in it. We may therefore establish things in

ZF in the standard classical way, knowing that they are perfectly acceptable from a paraconsistent perspective." (2006, pp.257-258)

So, on the model-theoretic strategy for paraconsistent set theory, we demonstrate that there exists a model of both *ZF* and our paraconsistent set theory, *M*, and, on that basis, we conclude that everything true on *ZF* is true on the universe of sets, materially construed by *M*. The advantage this has over the material strategy, where we try to prove things directly in *M* without a proper conditional, is that, in addition to the small number of results obtainable in *M*, we gain all the extra results about the universe of sets provided by *ZF*.

There are a number of problems with this approach. One is that we are not, in general, entitled to conclude on the basis of the existence of a model of two theories, in this case *ZF* and *M*, that everything true on one, *ZF*, is true on the other, *M*. We are not, for example, entitled to conclude that the Continuum Hypothesis is true according to *ZFC*, on the basis of Gödel's results that there are models of *ZFC* + CH. Or, as Weir (2004, p.398) puts the objection, the existence of a model of *ZFU* (*ZF* with urelements) in which Scotland wins the world cup does not entail that *ZFU* shows Scotland will win the world cup. So it doesn't seem that Priest is entitled to conclude that *M* contains all the truths of *ZF*, and, if not, this undermines the model-theoretic strategy.

A second problem is that the model-theoretic strategy employs the wrong logic. As Priest himself says "If the paraconsistent strategy for set theory is to be anything more than an intellectual exercise, the underlying logic used must, in some sense, be the right one for reasoning about sets." (2006, p.258) Priest is not a logical pluralist: he has argued in a number of places, correctly in my view, that logic is universal, and validity is to be defined as preservation of truth in *all* situations (for example, in his (2008) and Chapter 12 of his (2005). He has also, as discussed in Section 3.0, defended as the correct logic, *LP* with the addition of a relevant conditional. If he is right, this is the correct logic for reasoning in any situation and, in particular, it is the correct logic for reasoning about the sets: all its inferences are truth-preserving in set-theory, as everywhere else, and we ought to expect the relevant strategy to be executable in this logic. But this is not the logic employed on the model-theoretic strategy,

and so, on the model-theoretic strategy, the underlying logic is not the right one for reasoning about sets.

Priest might respond that *LP* without a conditional is a sub-logic of the one he takes to be correct (*LP* with the addition of a suitable conditional). This is true, but it's not clear that it helps. The correct logic, we're supposing, is *LP* plus a conditional and, if so, all the valid inferences of that logic should be valid when reasoning about sets. So, if Priest is right about the correct logic, we should be able to reason about sets using all of it, including with a proper conditional, rather than being restricted to a conditional-free sub-logic.

Another problem for this approach is that it is highly incomplete. As far as we know, we cannot establish directly in *M* anything like enough results for an adequate account of the universe of sets. Setting aside the objections above, the model-theoretic strategy promises to add to these results all the truths establishable in *ZF*. But this is not enough. Inconsistent naïve set theory goes far beyond what can be characterised in *ZF*: importantly, it contains universe-sized sets (indeed, the set of such sets is itself universe-sized). But the model-theoretic strategy has almost nothing to say about this vast part of the universe because, if something cannot be established in *ZF* (and then appropriated by *M*), it must be established in *M* directly, without a conditional satisfying *modus ponens*. So given the paucity of results establishable in this way, a crucial, and exceedingly large, part of the universe of sets is woefully under-described.

3.5.4 Paraconsistent Metatheory

Priest considers an objection to dialetheism that dialetheism requires a metatheory in which its semantics and proof systems are characterised and in which important results such as soundness and completeness can be proved. The only available metatheory, so the objection goes, is classical logic and so paraconsistent logic is incorrect (Priest, 2006, p.258).

His response to the objection is to deny that a classical metatheory is the only available. Instead, we should take paraconsistent set theory to be our metatheory. He points out that whether model theory can be adequately characterised on the relevant strategy for set theory, as outlined above, is an

open question (though he wrote this before Weber's results had been established, so may now be more hopeful on this point). Instead, he suggests, that we employ the model-theoretic strategy, carry out our metatheory in *ZF*, then simply appropriate the results into our paraconsistent set theory, *M*.

The objection he considers is not, it seems to me, characterised as well as it might be. Paraconsistent logic is not a logic but a family of logics, unified by the fact that, in those logics, the principle of explosion is invalid. The demand merely that paraconsistent logics be given paraconsistent metatheories, in light of this, is too weak on its own. For example, it would be far too weak a constraint on non-paraconsistent logics that they must have as their metatheory a non-paraconsistent logic, since this would allow paracomplete views to employ a classical metatheory with impunity. Rather, the constraint ought to be that a paraconsistent logic should have its metatheory carried out *in the very same logic*. A logic pluralist might reject this constraint, claiming that different logics can be true of different domains, none is uniquely correct and, perhaps, two such differing domains, requiring different logics, are the object theory and the metatheory. They would, of course, have to independently motivate this difference in treatment in these particular cases. But, as above, Priest is not a logical pluralist and so, it seems to me, has no reason to reject the constraint. Whatever the correct logic, it is the correct logic in every situation, including the logic's own metatheory. In light of this, the model-theoretic strategy cannot be employed in giving Priest's paraconsistent metatheory. This is, of course, much the same objection raised against the model-theoretic strategy in the context of set-theory itself.

Another problem with the model-theoretic strategy, which is particular to its being employed metatheoretically, concerns the definition of validity. Priest defines validity here (simplifying to the one-premise case) such that and inference from α to β is valid just if

Val for every interpretation, I ($I \Vdash \alpha \supset I \Vdash \beta$)

The '$\supset$' here is the material conditional which, as noted, does not detach. In fact, calling '$\supset$' the material *conditional* is slightly contentious, since, as Priest himself has noted, "[a]ny conditional worth its salt,$\rightarrow$, should satisfy the *modus*

ponens principle: {α, α → β}⊨β. This is, indeed, analytically part of what implication is."(2006, p.83) So, by Priest's own light, '⊃' is not a conditional at all. Moreover, he also notes "[f]or exactly the same reason, its necessitation is not the entailment connective either. Let α be some paradoxical sentence, which is not only a dialetheia, but necessarily so (such as the liar sentence). Then, for any β, $\neg\alpha \vee \beta$ is not only true, but necessarily so. This helps not a whit in inferring β from α." (2006, p.83)

This seems quite right, and so it is peculiar, and problematic, that Priest is now characterising validity using this very 'conditional', of which he has previously said "… dialetheism disposes, once and for all" (2006, p.83). Priest anticipates an objection to this, according to which defining validity in this way "deprives the notion of validity of its punch" (2006, p.259). He claims that it does not, and gives two reasons why. The first is that, though *modus ponens* is strictly invalid, it is perfectly usable, if the situation is consistent. So we may detach the conclusion of a valid argument, so long as the premises are true and consistent (which, in this context, means that we don't have both $l \Vdash \alpha$ and $l \nVdash \alpha$, which I'll call 'externally inconsistent' premises, which is different from the 'internally inconsistent' $l \Vdash \alpha$ and $l \Vdash \neg\alpha$). The second is that, whether characterised relevantly or materially, the conditional in *Val* is simply a true or false statement, whereas inference is an action, to which no number of true statements is equivalent (this, he says, is the moral of Lewis Carroll's dialogue between Achilles and the Tortoise (1895)). Still, the statements may ground the action, but there is no reason, Priest thinks, that statements containing material conditionals may not do this every bit as much as those containing relevant conditionals; it is just that relevant conditionals always ground the action, but material conditionals only sometimes.

It's unclear exactly what is the content of the objection to which Priest is responding, or what exactly 'depriving validity of its punch' amounts to. But, it seems to me, there are two problems with this account, neither of which is addressed either by Priest's references to his views on classical recapture, or to the moral of Carroll's dialogue. The first is that, even granting Priest's account of classical recapture and that we can happily detach the consequent in externally consistent cases, we still can't do so in externally inconsistent cases.

The effect of this is that we simply cannot reason from externally inconsistent premises. If the premises are such that we have $I \Vdash a$ and $I \nVdash a$, then we can never obtain the truth conclusion, whatever it is. This is a significant limitation for a dialetheist. There is nothing illegitimate, by the dialetheists lights, about such contradictions, and as such, we should expect to be able to reason about them, and to discuss what truths follow from them (as some things clearly must: to take the most trivial example, presumably they follow from themselves). But we cannot do this, and so the material definition of validity is inadequate.

A second problem, alluded to above, is that detachment is analytically part of what validity *is*, and so *Val* cannot be a definition of validity, if it is understood materially. Priest accepts that detachment is a "*sine qua non* of any implication connective" (2006, p.86), in which he includes the entailment connective expressing what follows logically from what. But the definition of validity simply tells us under what conditions the conclusion of an argument follows logically from its premises. So, if it is a *sine qua non* of any implication connective, including entailment, that it detach, it should be a *sine qua non* of validity that the conclusion of a valid argument can be detached when the premises are true. At the very least we need some account making sense of this asymmetry. I am doubtful that any plausible account vindicating the difference can be given, and in its absence, we should conclude that non-detachable validity is not validity at all.

A related problem is that each of the logical connectives, $\wedge$, $\vee$, $\neg$, $\rightarrow$, plausibly, obeys principles which are, at least in part, constitutive of the meaning of the connectives in question. For example, it is part of what conjunction *is* that, whenever we have $\{A, B\}$ true, we also have the truth of $A \wedge B$. But if the truth of the premises of a valid argument does not, in general, allow us to obtain the truth of its conclusion, this is no longer so. There are infinitely many A's and B's such that we have $\{A, B\}$, but cannot conclude that $A \wedge B$ is true. So, if validity doesn't always deliver the truth of the conclusion, given the truth of the premises of a valid argument, then not only don't we have a notion of validity, we don't have a proper conjunction, or any other of the connectives, either.

In sum, the model-theoretic strategy faces decisive objections, both as an approach to set theory and, more specifically, as a metatheory for dialetheism.

The relevant strategy is more promising. Set theory, carried out in a relevant logic, is still in its infancy, and it remains to be seen whether it can be developed into a fully-fledged account of the sets comparable with set-theory classically-understood, but Weber's results give us some reason to be hopeful.

3.6 Chapter Conclusion

This chapter was about metatheoretically paraconsistent dialetheism, especially the version defended by Graham Priest in his (2006). He motivates his view in a negative way by attempting to rule-out the alternatives to dialetheism. The most important arguments against opposing accounts of the paradoxes was the problem of revenge, though there were other arguments too (importantly, those concerning set theory). One issue, of crucial importance for revenge, is the metatheory of the view. This is, on the version of dialetheism under discussion, paraconsistent (indeed, inconsistent) set theory. Priest gives a few strategies by which this might be carried out. I have argued against two of these, and suggested that the most promising is the so-called 'relevant strategy'.

Chapter 4: Getting Revenge on Metatheoretically Paraconsistent Dialetheism

4.0 Introduction

The previous chapter characterised a thoroughgoing version of dialetheism whose metatheory is paraconsistent (indeed, inconsistent) set theory. The only philosopher who has a thoroughly developed version of this view is, to my knowledge, Graham Priest, and so the chapter focused on the view as developed by Priest (principally in his book, *In Contradiction* (2006)).

The purpose of this chapter is to show that the sort of thoroughgoing dialetheism defended by Priest is subject to a revenge problem. Showing that this is the case is not easy. As mentioned in the previous chapter the fact that Priest's object language gives its own, inconsistent, metatheory, makes it very resilient against standard strategies for revenge. As I said, the attempt to get formal revenge on the theory i.e. to show that we can construct notions in its metatheory which would trivialise in the object theory, is collapsed into the attempt to demonstrate the triviality of the theory as a whole. The prospects for this do not look good.

The possibility of developing what I call an 'informal revenge' problem, however, is still a live one, and this is the strategy pursued in this chapter. We show that a view suffers an informal revenge problem by demonstrating that there are certain notions which appear in dialetheists' informal characterisation of their view which cannot be expressed in the dialetheist's theory. We prefer notions which appear in the dialetheist's characterisation of their view, since the dialetheist is thereby committed to the legitimacy of the notion, and cannot simply dismiss it as incoherent.

A number of attempts have been made at showing dialetheist suffer expressive limitations, and I discuss some in this chapter, with a particular focus on the one which seems most closely-related to my own: the so-called 'just false' problem.

First, I discuss attempts to demonstrate that dialetheists are unable to express each of a cluster of notions involved in the characterisation of 'normal'

sentences: that is, sentences, or sets thereof, which are not of the kind dialetheists treat inconsistently. I also discuss some responses to these problems, especially Priest's ((2006), (2010a)).

Then, I develop my own revenge problem, which concerns the notions 'invalid' and 'just invalid'. First, I show that one feature of the dialetheist's paraconsistent metatheory is that it commits them to dialetheism, not just about truth, but about validity: that is some inferences will be both valid and invalid. I will argue that the extent of the overlap between the valid and the invalid inference principles is so great as to rob dialetheists of the ability to express invalidity. The exact extent of the overlap may well depend on the details of the dialetheist's view: in particular, the behaviour of the truth, and of the conditional. On some views, the conclusion that every inference principle is invalid is easy to establish; not so on others. I discuss these in turn.

Once the view that some contradictions are true was introduced, it became important to be able to express the status of the other 'normal' sentences, for example, the 'just false' sentences. Similarly, if validity is shown to be inconsistent, and some inference principles both valid and invalid, we ought to be able to express the status of the principles which are 'just invalid'. But, as I point out, on the obvious ways of characterising the notion, 'just invalid' is equivalent to invalid. Moreover, a necessary condition on characterising a notion of 'just invalid' for which this is not the case is characterising a notion of 'just false' which behaves consistently. The revenge problem, then, is that both 'invalid' and 'just invalid' overlap so thoroughly with notions like 'valid' and 'inference principle' that they are deprived of their meaning. For versions of dialetheism for which the stronger result can be established that every inference principle whatever is invalid (which may yet be all available views), then 'inference principle', 'invalid' and 'just invalid' are logically equivalent notions. I consider some possible responses to this problem on behalf of dialetheism and argue that they are inadequate.

4.1 Revenge, Just False and Non-Dialetheia

Dialetheists admit that there exist true contradictions but, as non-trivialists, they take these dialetheia to be the exception: most sentences are not like this. But dialetheists seem to run into expressive difficulties when trying to classify the 'normal' sentences as such. The reason for this is that the cluster of notions one would wish to employ in these characterisations turn out to be inconsistent. This has been developed into an objection to dialetheism by a number of philosophers (for example, Batens (1990), Parsons (1990), Littman and Simmons (2004), but the objection is best presented, it seems to me, by Shapiro (2004), and so I draw on Shapiro's characterisation of the problem in what follows.

Shapiro argues that dialetheists cannot express the notions 'just true' (though he calls this 'simple truth'), 'just false' (though he calls this 'simple falsity'), 'non-dialetheia', consistency and, relatedly, they cannot express disagreement. The reasons for each are, in essence, the same: it seems part of the meaning of these notions that they behave consistently, or, at the least, not too inconsistently. In this chapter, I discuss two of these notions, 'just false' and 'non-dialetheia', the reasons we might think they are inexpressible for dialetheists and some dialetheist responses to the problems. I discuss these two only since they relate most closely to my own revenge problem, and so best help illustrate it, and because the arguments concerning the others are similar enough that these should be sufficient to illustrate the general problem.

One very simple way of characterising dialetheism (of the non-trivialist variety) is as the view that some statements are both true and false, some are just true and some are just false. Claims of this kind appear with great frequency in descriptions of dialetheism, by dialetheists as well as others. So dialetheists ought to be able to express the notions involved in such claims, especially 'just true' and 'just false'. As I have said, I focus here on the latter notion (though much the same things are true of the former). We take the obvious definition of 'just false' as 'false and not true', as Priest does (for example in his (2010a), p.137, though he endorses this in a number of other places too).

We can see that the notion of 'just false', defined in the obvious way, is inconsistent by considering the following sentence, Σ, equivalent to:

$F\langle\Sigma\rangle \wedge \neg T\langle\Sigma\rangle$

Intuitively, this sentence says of itself that it is just false. We can derive contradiction from Σ in much the same way as with the strengthened liar sentence:

(1) $T\langle\Sigma\rangle$	(suppose for reductio)
(2) Σ	(1, T-scheme)
(3) $F\langle\Sigma\rangle \wedge \neg T\langle\Sigma\rangle$	(2, substituting equivalents)
(4) $\neg T\langle\Sigma\rangle$	(3, $\wedge$-elimination)
(5) $\neg T\langle\Sigma\rangle$	(1, 4, reductio)
(6) $F\langle\Sigma\rangle$	(5, exhaustion)
(7) $F\langle\Sigma\rangle \wedge \neg T\langle\Sigma\rangle$	(5, 6, $\wedge$-introduction)
(8) Σ	(7, substituting equivalents)
(9) $T\langle\Sigma\rangle$	(8, T-scheme)
(10) $T\langle\Sigma\rangle \wedge (F\langle\Sigma\rangle \wedge \neg T\langle\Sigma\rangle)$	(7, 9, $\wedge$-introduction)

So Σ is both true and just false, and so the notion 'just false', defined in this way, is inconsistent. In fact, matters are simpler than this. Though it is not Priest's view, a number of other philosophers (for example, Field (2008) and Beall (2009)) have endorsed the view that the T-scheme should be fully transparent and, hence, contraposable; so (on the assumption that falsity is defined as truth of negation) $F\langle\alpha\rangle \leftrightarrow \neg T\langle\alpha\rangle$ holds. On this view, take a sentence, α, and suppose it is false, so we have $F\langle\alpha\rangle$. Given the properties of the T-scheme just mentioned (in particular, what Priest calls 'exclusion'), we have $\neg T\langle\alpha\rangle$, and so $F\langle\alpha\rangle \wedge \neg T\langle\alpha\rangle$. So, on the assumption that α is false, α is *just* false. Since, obviously, the converse also follows, 'false' and 'just false' are equivalent, on this view. So every falsehood is just false, including every dialetheia. So, paraphrasing Shapiro (2004, p.342), saying that something is just false says no more than that it is false: the 'just' comes for free and adds nothing.

Since, on this version of dialetheism, 'just false' makes no distinction between dialetheia and the non-dialetheia we wish to pick out, it fails to capture what

we want when we utter expressions like ‘Some sentences are both true and false, some are just true and some are just false’. For all the distinction drawn, we may as well have said ‘Some sentences are both true and false, some are true and some are false’, which is not at all what was intended.

Priest responds on behalf of this variety of dialetheism that “…this objection is entirely question-begging. If one does subscribe to the exclusion scheme, there *is* no distinction—nor, therefore any work to be done to draw it.” (2006, pp.292-293) I disagree that the objection is question-begging. There certainly seems to be a difference between ‘false’ and ‘just false’, captured by the difference between sentences like ‘Some sentences are false, but some are just false’, on the one hand, and ‘Some sentences are false, but some are false’, on the other. These, and sentences like them, again, are uttered by dialetheists frequently, and it seems to me that we clearly understand what is intended. To pick an arbitrary example, Priest and Berto (1998) characterise an historical form of dialetheism as follows:

“In ancient Indian logic/metaphysics, there were standardly four possibilities to be considered on any statement at issue: that it is true (only), false (only), neither true nor false, or both true and false. Buddhist logicians sometimes added a fifth possibility: none of these.”

For another example, just before JC Beall defends the view that ‘true’ and ‘just true’ are equivalent (since his metatheory is consistent, I leave a proper discussion of his view until Chapter 5), he says:

“Rational dialetheists maintain that some (actually, many) [truths] are *just true*; they reject that all or even most claims are gluts. Indeed, on my account, it is only the spandrels of [truth] (or related notions) that are gluts; the rest are ‘just true’.” (2009, p.48)

The ‘only’ used in the first passage adds something to the characterisation of the view described above (why else would the authors have written it?) Similarly, though perhaps Beall adds ‘scare quotes’ to indicate that the notion he is discussing is problematic, I suspect most readers understand exactly what these sentences mean, and so grasp a difference between ‘true’ and ‘just true’

(and, correspondingly between ‘false’ and ‘just false’), which can’t simply be dismissed.

If it were a straightforward consequence of dialetheism generally that this distinction could not be accounted for, this would be a problem for dialetheism, and it would not be question-begging to says so. But the equivalence of falsity and just-falsity is not an immediate consequence of dialetheism, and so the charge of question-begging seems to me to have even less plausibility.

Though many philosophers have endorsed the account of the T-scheme on which the preceding remarks are based, as I have said, this is not Priest’s view. Priest rejects the principle of exclusion according to which falsity entails untruth: $F\langle\alpha\rangle \rightarrow \neg T\langle\alpha\rangle$. This means that the situation for Priest, with regards to the ‘just false’ problem, is different from that sketched immediately above. In particular, since falsity does not entails untruth, we cannot in general show that an arbitrary falsehood is thereby untrue and, hence, just false.

The notion of just false, though, is still inconsistent for Priest, since he still accepts the proof above. Moreover, we can show, since Priest accepts exhaustion, $\neg T\langle\alpha\rangle \rightarrow F\langle\alpha\rangle$, that any sentence, α, which is untrue, is thereby false and, hence, just false. So the strengthened liar, λ, equivalent to $\neg T\langle\lambda\rangle$, since it can be shown to be both true and untrue, is both true and just false. The same holds for every untrue dialetheia, of which there are infinitely many.

Notions like just false certainly seem, on the face of it, to have as part of their meaning that they behave consistently. Still, it might be thought, it is unfair to expect these notions to behave consistently for a dialetheist, since they are semantic notions which can be employed in the construction of liar sentences, which are the very sorts of thing dialetheists urge us to treat inconsistently. But, Shapiro claims (2004, p.342), the point of introducing the notion of just false in the first place was to distinguish falsehoods from at least paradigmatic dialetheia, which the strengthened liar is. So, even if we don’t demand the notion behave consistently, if the strengthened liar, as well as infinitely many other untrue dialetheia, are just false, we might think the notion too inconsistent to be of use.

In response to this, Priest says “I am not sure what the first place was in this case, nor, therefore, what its point was. But since $F\langle\alpha\rangle$ and $F\langle\alpha\rangle \wedge \neg T\langle\alpha\rangle$ are not logically equivalent, there is a distinction between being false and being simply false. The fact that some sentences (be they paradigm dialetheias or anything else) may be in both camps is just one of those contradictory facts of life that populate the dialetheic landscape. [some sentences are]… simply false—and…[some are]…true as well.”(2006, p.294)

So Priest’s response to the problem is to simply accept the inconsistency of ‘just false’ and deny that this prevents it from meaning just false, regardless of its inconsistency in the case of ‘paradigm’ dialetheia, because we still have a distinction (because of the failure of exclusion) between being false and being just false.

This response, in a way, is an admission of an expressive limitation, as Priest has said himself elsewhere, he “cannot assert that something is false-only if this is required to exclude things that are true as well.”(2010a, p.136) He does not think this constitutes a revenge problem, however. The reason is that the consistent theories which he argues are subject to revenge are completely unable to express a key notion in their theory: they have no way to express ‘not true’, ‘not determinately true’, ‘defective’, or whatever the notion involved in the revenge problem happens to be. His position, he thinks, is not like this. He cannot express ‘just false’ where this is understood consistently, but he can express it inconsistently: “[t]he dialetheist about the paradoxes does have a way of expressing that something is false only – in the very words ‘false and not true’.”(2010a, p.137)

This response is adequate only if the notion expressed when dialetheists use the expression ‘just false’ still means ‘just false’, despite its inconsistency (especially in paradigm cases like the strengthened liar). I have sympathy with Shapiro’s claim that it does not. On the other hand, it is difficult to see what argument can be given to this effect whose premises a dialetheist like Priest would be happy to accept. This is not, of course, to concede that Priest is right; rather the situation seems to have reached a stalemate. It is worth re-iterating, though, that this response from Priest is only available to him because of his

rejection of exclusion. If one accepts the equivalency of falsity with untruth, the distinction between falsity and just-falsity is collapsed.

A closely related problem concerns the notion 'non-dialetheia'. Shapiro points out (2004, pp.342-343), attributing the original complaint to Parsons (1990, pp.345-346 n. 10), that there are a number of ways of expressing that something, α, is a dialetheia. We might say, though this may not exhaust the possibilities, $\alpha \wedge \neg\alpha$, or $T\langle\alpha\rangle \wedge F\langle\alpha\rangle$, or perhaps $T\langle\alpha\rangle \wedge \neg T\langle\alpha\rangle$. Corresponding to each of these is a way of expressing what it is to be a non-dialetheia: $\neg(\alpha \wedge \neg\alpha)$, $\neg(T\langle\alpha\rangle \wedge F\langle\alpha\rangle)$ and $\neg(T\langle\alpha\rangle \wedge \neg T\langle\alpha\rangle)$, respectively. Each of these is inconsistent, in the sense that each holds of some dialetheia (so some dialetheia are non-dialetheia, on each of these definitions). The first, $\neg(\alpha \wedge \neg\alpha)$, is a dialetheic logical truth, and so holds for any sentence α, including every dialetheia. So this option is inadequate. The third option, $\neg(T\langle\alpha\rangle \wedge \neg T\langle\alpha\rangle)$, is also true of every sentence, α, as Priest has shown (2006, pp.90-91). The second option, $\neg(T\langle\alpha\rangle \wedge F\langle\alpha\rangle)$, differs only if exclusion fails. If both exclusion and exhaustion hold, then falsity and untruth are equivalent and, hence, $\neg(T\langle\alpha\rangle \wedge \neg T\langle\alpha\rangle)$ and $\neg(T\langle\alpha\rangle \wedge F\langle\alpha\rangle)$ are equivalent. This would mean the second option is also true of any α. Since Priest rejects exclusion, things are different for him.

The situation is similar to the one described above for 'just false'. Since it does not follow from the falsity of some sentence, α, that α is untrue, it does not follow from the falsity of α that α is not both true and false. So, for Priest, this definition of non-dialetheia, though inconsistent (some dialetheia are non-dialetheia), is not vacuous (not all dialetheia are non-dialetheia).

Unsurprisingly, Priest chooses the option which is not vacuous: *a* is a dialetheia if $T\langle\alpha\rangle \wedge F\langle\alpha\rangle$ and so a non-dialetheia if $\neg(T\langle\alpha\rangle \wedge F\langle\alpha\rangle)$. Though the definition is not vacuous, Shapiro still thinks it is problematic. Any sentence which is untrue is thereby *not* both true and false, and so not a dialetheia, on this definition. This includes paradigm dialetheia like the strengthened liar. He says:

"It is surely reasonable to demand that our definitions be non-vacuous. If we have to say that *every* sentence is a non-dialetheia, then the notion is useless. And the notion is all but useless if we have to say that every untruth is also a non-dialetheia, including the Original Liar and nearly all the dialetheias we run across in the course of thinking about this stuff—every non-truth and every non-falsehood are also non-dialetheias. I would have thought that a dialetheist, like Priest, would *deny* that the Original Liar is a non-dialetheia, rather than asserting that it is one." (2004, p.344)

Priest's response is to down-play the overlap between the dialetheia and the non-dialetheia:

"Shapiro overplays his hand here. There is indeed an overlap in the categories of being and not being a dialetheia. But the only denizen of the overlap we have on the table is ξ(¬ T<ξ>). And this is very special. By its particular properties, it is true and it is not true—and so not (true and false). But the same is not the case for any other of the standard paradoxes of self-reference (including the liar in the form 'this sentence is false'). (See the comments on chapter 1 in section 19.3) And I am quite happy to assert that ξ is a non-dialetheia, provided that I can add that it is as well." (2006, p.294)

The strengthened liar may be the only denizen of the overlap 'on the table', but it is certainly not the only denizen of the overlap there is. For one, once we have one dialetheia of this kind, it follows that there are, by similar constructions, infinitely many. So at least in terms of cardinality, there is not only one such contradiction. Still, a more charitable way of reading Priest is as claiming that the strengthened liar is the only *kind* (in some sense) of paradox which is both true and untrue. But this isn't right either. At least some of the other paradoxes of self-reference can be 'strengthened' in a similar way to the liar.

For example, Grelling's paradox is treated dialetheically by Priest. Grelling's paradox arises from the predicate 'heterological', which is defined as being true of a predicate when that predicate does not apply to itself. For example,

‘monosyllabic’ is heterological, since it is not monosyllabic. The contradiction is produced by applying the predicate to itself: that is, by asking whether ‘heterological’ is heterological. But we can define heterologicality differently to obtain a strengthened version of the paradox: define a predicate to be heterological when the result of applying it to itself is, not false, but untrue. The dialetheia generated by this version of the paradox is both true and untrue and, hence, a non-dialetheia.

A little more precisely, if we introduce our heterologicality predicate, *Het*, defined such that it holds of a predicate just if it is untrue that the predicate applies to itself, then we can derive contradiction as follows:

(1) $T\langle Het\langle Het\rangle\rangle$	(Suppose for *reductio*)
(2) $Het\langle Het\rangle$	(1, T-scheme)
(3) $\neg T\langle Het\langle Het\rangle\rangle$	(Definition of *Het*)
(4) $\neg T\langle Het\langle Het\rangle\rangle$	(1, 3, *reductio*)
(5) $Het\langle Het\rangle$	(4, definition of *Het*)
(6) $T\langle Het\langle Het\rangle\rangle$	(5, T-scheme)

Another example is Russell’s paradox, which can be strengthened by introducing a satisfaction predicate, along with the naïve principles for satisfaction. We take the Russell set to be $\{x: \neg Sat(x, \langle v \in v\rangle\}$ and prove a contradiction as follows:

(1) $R \in R$	(Suppose for *reductio*)
(2) $\neg Sat(R, \langle v \in v\rangle)$	(1, comprehension)
(3) $Sat(R, \langle v \in v\rangle)$	(2, exhaustion)
(4) $R \notin R$	(3, definition of satisfaction)
(5) $R \notin R$	(1, 4, *reductio*)
(6) $Sat(R, \langle v \in v\rangle)$	(comprehension, contraposition, DNE)
(7) $R \in R$	(6, definition of satisfaction)
(8) $\neg Sat(R, \langle v \in v\rangle)$	(7, comprehension)

Since for x to satisfy φ is for φ to be true of x, it is both true and untrue of this Russell set that it is self-membered, and we have another denizen of the overlap.[5]

It is not clear to me at the moment how many of the commonly discussed paradoxes of self-reference (others being Berry's, Quine's or Yablo's) can be strengthened in this way, though it strikes me as likely that they can. Even if they cannot, the examples given seem like enough to undermine Priest's claim that the strengthened liar is unusual or special, among the self-referential paradoxes, in being both true and untrue. This lends further strength to Shapiro's claim that the overlap is too great for the notion of 'non-dialetheia' to be of much use. This is especially so if a similar treatment can be extended to other paradoxes.

On the other hand, the notion is still not vacuous: not every dialetheia is a non-dialetheia. In particular, Priest may still think that the fact that a significant number of the self-referential paradoxes, especially those characterised using falsehood rather than untruth, are both true and false but *not* untrue, is sufficient for the term 'non-dialetheia' to preserve its meaning and usefulness.

My sympathies are with Shapiro here, that non-dialetheia ought to behave consistently, or at least, less inconsistently than it does, for it to mean what's needed. But as with the just false problem, it is difficult to see how, if Priest takes the view described above, one can argue for this conclusion from premises Priest is likely to accept. Again, though, this is only because Priest rejects exclusion: if this principle is accepted, all of the obvious definitions of non-dialetheia collapse into vacuity.

The problems discussed, of the apparent expressive limitations of dialetheism, arise from the obvious definitions of 'just false' and 'non-dialetheia', respectively. One avenue of response is to define a new, unobvious, definition of 'just false' and, correspondingly, of 'non-dialetheia', which behaves consistently.

[5] My thanks to Alan Weir for the proof just given.

The most obvious problem facing such attempts is the threat of revenge paradoxes employing the newly defined notion. So, for any newly-defined notion *just false*, we construct a sentence:

(*) The sentence marked (*) is *just false*

From this sentence, little is required to derive the conclusion that (*) is both true and *just false*. This obstacle must be avoided if the notion to do its job, which is no easy task.

Still, attempts have been made to solve the problem in a way which avoids this. Some options which have been considered are a form of primitive exclusion (Berto, 2014), a form of negation called 'arrow falsum', as well as non-logical 'shriek rules' (see Beall (2013) for a discussion of the second two).

I don't discuss these options here for a few reasons. The main reason is that each of the attempts at solving this problem with a new definition of 'just false' has been carried out in a setting where the metatheory is taken to be consistent. The views under discussion in this chapter are ones which take the metatheory to be inconsistent. It is not obvious how these solutions behave in this different setting, and it seems more appropriate to discuss the details of them in Chapter 6, which concerns metatheoretically consistent dialetheism.

A second reason is that to take devices like arrow-falsum, shriek rules or primitive exclusion as putative solutions to the just false problem is to conflate the just false problem with the, admittedly related, problem of exclusion. This conflation appears frequently in the literature, but it occurs importantly in Beall's response, based on shriek rules, to this problem. So I explain in detail the differences between just false and exclusion when I turn to potential revenge problems for Beall's version of dialetheism in Chapter 6.

I do, however, discuss Berto's primitive exclusion device, briefly, in this chapter, since one might think something like this device could help with my own revenge problem for dialetheism.

So, although it may be that some new, consistent notion of ‘just false’ can be defined for metatheoretically inconsistent dialetheism, no such definition has, to my knowledge, been offered. The principal reason for this, I suspect, is that if such notions are strong enough to capture anything like ‘just false’, a revenge liar will demonstrate them to be inconsistent, defeating their purpose. So Priest, the most prominent proponent of this version of dialetheism, simply settles for the original definitions and accepts their inconsistency. This seems to me, at least dialectically, the strongest option for proponents of this sort of view to take.

Another strategy one might wish to pursue here is to take ‘just false’ to be some kind of pragmatic device, perhaps a Gricean conversational implicature or some such. So, if a dialetheist says of some sentence that it is just false, this somehow pragmatically indicates that the sentence is not also true. An account of how exactly this works, and how it addresses the just false problem in particular would have to be given. One initial reason to be sceptical of this approach is that, by employing this strategy, the dialetheist has accepted that there is no expression whose actual content captures that something is just false, since, if there were such an expression, there would be no need to invoke pragmatic devices. This being the case, we might wonder exactly what kind of strange, ineffable fact is being pragmatically alluded to, but whose content cannot be captured by any expression. As Ramsey famously quipped, though about Wittgenstein’s *Tractatus*, “What we can’t say we can’t say, and we can’t whistle it either.” (1990) Whether dialetheists taking this approach would really be trying to whistle something that can’t be said will depend on the details of the approach yet to be provided. But it seems likely that in giving this account, the dialetheist will have to assert some statement indicating what the pragmatic device is supposed to be capturing. Some independent reason for why we should not simple take this statement to capture the theory’s account of ‘just false’, employ it in the construction of a liar sentence, and so demonstrate that the account is really inconsistent, will have to be given. If the only reason this is debarred is that it is necessary to avoid inconsistency, then the account is ad hoc.

Another, perhaps more decisive objection is given by Shapiro (2004, p.339), who points out that pragmatic devices of this kind cannot be embedded in the antecedent of a conditional, or within the scope of a negation, or in a hypothesis. Since we clearly can embed expressions containing 'just false', as well as negate them and form hypotheses containing them, the just false problem cannot be solved this way.

4.2 Inconsistent Validity and Revenge

Dialetheism which takes classical *ZF* set theory as its metatheory, of course, allows some sentences to be both true and false: that is, some sentences are assigned both 1 and 0 in the same model. But this inconsistency infects the metatheory no more than my views become inconsistent when I report that, according to Graham Priest, some sentences are both true and false. But if the metatheory is inconsistent, we should expect the assignments of truth-values to sentences themselves to become inconsistent: that is, we should expect some sentences both to be assigned 1 and not assigned 1 (in a model), and some sentences both to be assigned 0 and not assigned 0 (in a model). For the situation to be otherwise would be extremely peculiar; we should expect the inconsistency of the metatheory to make itself manifest.

As I show in this section, we do indeed find that metatheoretically inconsistent dialetheism makes inconsistent assignments of truth values to sentences. One significant effect this has is to make, not just truth, but *validity* dialetheic. This fact is significant and interesting on its own, but it has greater significance because it leads to a revenge problem: we can show that so many inference principles (perhaps all of them) are invalid that dialetheism becomes unable to express invalidity. This is a very serious expressive limitation, since there are crucial semantic facts about the theory, for example the invalidity of the principle of explosion ($A, \neg A \vDash B$), which the dialetheist must use the notion to express.

The ease with which, and the extent to which, it may be shown that inference principles are invalid on this view varies with respect to a handful of factors,

especially the behaviour of the truth-predicate. So I take a number of options in turn and show the extent of the overlap between validity and invalidity for each option. On the first two options, it is straightforward to show the invalidity of every inference principle there is. On the second two options, because of the behaviour of Priest truth predicate (in particular, again, the failure of exclusion), things are more difficult. I begin by showing that every inference principle capable of taking an atomic sentence as its conclusion is invalid. I then extend this, using a result by Richard Heck (2013), to every inference principle whose conclusion contains only extensional connectives.

Next, I address principles whose conclusion contains a conditional. I show, in a piecemeal way, that a number of these are invalid. I do not, at the present, have a general result that every such principle is invalid, though I suspect there is one to be had. I discuss a number of avenues of response to this problem, and argue against them. Throughout most of this chapter, I assume that the dialetheic metatheory under discussion is carried out in paraconsistent set theory, where the underlying logic is *LP* with the addition of a suitable, relevant conditional. I first discuss the view which takes truth to be transparent, and so which takes the biconditional in the T-scheme to contrapose. I don't take a view on whether the conditional added to *LP* ought itself to contrapose, since if it does not, the same effect may be achieved (as regards the T-scheme) by stipulating, in addition to $T\langle\alpha\rangle \leftrightarrow \alpha$, $\neg T\langle\alpha\rangle \leftrightarrow \neg\alpha$. I conclude the chapter with a brief discussion of the prospects of developing my revenge problem on Priest's model-theoretic strategy. I give arguments which seem to show that there must be metatheoretically inconsistent assignments of truth values to sentences (some both get 1 and don't), but point out the difficulties involved in showing how these might invalidate valid inference principles.

Validity is understood throughout as necessary preservation of truth, as defined by *Val*, such that $A \vDash B$ iff

(*Val*) for every interpretation *I*, $(I \Vdash A) \rightarrow (I \Vdash B)$

I assume throughout, as Priest accepts (2006, p.87), that the conditional validates the principle $A \wedge \neg B \vDash \neg(A \rightarrow B)$ and so any instance such that A relates to 1 while B fails to relate to 1 will invalidate $A \vDash B$.

4.2.1 Inconsistent Validity with Transparent Truth

Though it is not Priest's view that truth is transparent, a number of philosophers have found this view attractive. If the adoption of metatheoretically paraconsistent dialetheism were to become more widespread, it seems reasonable to suppose that a number of dialetheists would want the T-scheme to behave in this way. In fact, this is the easiest sort of view for which we can demonstrate the inconsistency of validity.

Defenders of this view accept the T-scheme in full generality. They also accept both of:

(*Exclusion*) $T\langle\neg\alpha\rangle \rightarrow \neg T\langle\alpha\rangle$

(*Exhaustion*) $\neg T\langle\alpha\rangle \rightarrow T\langle\neg\alpha\rangle$

One would expect them to accept similar principles regarding truth-in-a-model (TIAM) so, in Priest's relational semantics, for any interpretation ρ:

(*TIAM-Exclusion*) $\langle\neg\alpha\rangle\rho 1 \rightarrow \neg\langle\alpha\rangle\rho 1$

(*TIAM-Exhaustion*) $\neg\langle\alpha\rangle\rho 1 \rightarrow \langle\neg\alpha\rangle\rho 1$

In fact, Priest suggests that they may not need to endorse these principles. He says, discussing a claim from Shapiro about the exclusion principle, that "[o]ne should note, here, that we are talking about truth/falsity *simpliciter*, not truth/falsity in an interpretation. Truth/falsity in an interpretation is quite

different. Even if one endorses the exclusion schema, this does not entail that, for any given interpretation, ρ, $<\alpha>\rho 0 \rightarrow \neg<\alpha>\rho 1$." (2006, p.293)

If one accepts *Exclusion*, it is highly problematic to reject *TIAM-Exclusion*. To accept an asymmetry of this kind between truth/negation *simpliciter* and truth-in-a-model/$\neg$ is to make the semantics of the theory *ad hoc* and unnatural. After all, the semantics are meant to be characterising the features of actual negation, and, via truth-in-a-model, actual truth. But if *Exclusion* holds of truth, though the corresponding principle fails of truth in a model, then there is an important asymmetry between the real notions and the notions as they appear in the semantics. In fact, it doesn't seem unreasonable to say that '$\neg$' is *not* negation and 'truth-in-a-model' is not *truth*-in-a-model.

Shapiro argues that Priest's semantics are artificial, since he denies *TIAM-Exclusion*, and is thereby denied a homophonic semantics (2004, p.352). Priest responds by denying that a homophonic semantics is desirable, and claims that the objection amounts to nothing more than the question-begging insistence that dialetheic negation is not negation (2006, pp.294-295). Moreover, he thinks, once we admit dialetheia, homophonic semantics are unappealing since, if α is a dialetheia, $\neg\alpha$ holds, but then so does α. This is the same argument as that discussed in 3.5.1.2, and is still unpersuasive in this context. Having said that, if Priest is right to reject *Exclusion* in the case of truth, he ought also to reject *TIAM-Exclusion* for truth-in-a-model, for the same reason that those accepting *Exclusion* of truth ought to accept *TIAM-Exclusion* for truth-in-a-model: if the semantics are otherwise there is a problematic asymmetry between the notions occurring in the model theory and the real semantic notions we are trying to model. So one ought to accept *TIAM-Exclusion*/*TIAM-Exhaustion* if and only if one accepts the corresponding principle about truth *simpliciter* i.e. *Exclusion*/*Exhaustion*.

So, on the view under discussion, the dialetheist has accepted both *Exclusion* and *Exhaustion*, as well as their truth-in-a-model counterparts, *TIAM-Exclusion* and *TIAM-Exhaustion*. Given these principles, the invalidity of every inference

there is follows straightforwardly from the existence of a trivial model of the logic *LP*. In the trivial model, every atomic sentence receives the values 1 and 0. By *TIAM-Exclusion*, then, every atomic sentence receives 1 and fails to receive 1. By *TIAM-Exclusion, TIAM-Exhaustion*, and the recursive clauses for the extensional connectives, every conditional-free complex sentence receives 1 and fails to receive 1. We extend this to sentences containing a conditional by considering, in Priest's tertiary semantics, the instance of the relation $Rw_{\perp}w_{\perp}w_{\perp}$ (that is, the instance which relates the trivial model, or rather, in this case, the trivial world, to itself). This makes every conditional false at the trivial world and, by *TIAM-Exclusion*, untrue at that world. This makes every inference principle invalid since each has an instance, at the trivial world, at which the premises are true and the conclusion fails to be true at that world.

We can illustrate with a couple of examples. To take the simplest, consider

(*Reflexivity*) $R \vDash R$

Take some atomic sentence, φ, and substitute for R to obtain the following instance of *Reflexivity*

$\varphi \vDash \varphi$

Since φ both relates to 1 and fails to relate to 1 in the trivial model, in this instance of *Reflexivity*, the premise relates to 1 and the conclusion fails to. So *Reflexivity* is invalid. To take another example, ∧-introduction:

(∧*-Intro*) $A, B \vDash A \wedge B$

Take the conjunction of sentences $\varphi \wedge \psi$, which receives 1 and fails to receive 1 in the trivial model (as do its conjuncts). Substituting, we obtain the following instance of ∧*-Intro*:

$\varphi, \psi \vDash \varphi \wedge \psi$

Again, since φ, ψ and $\varphi \wedge \psi$ receive 1 and fail to receive 1 in the trivial model, the premises in this instance of $\wedge$-*Intro* receive 1, the conclusion fails to receive 1 and so the inference is invalid. This strategy invalidates every inference principle there is.

Though I have argued that dialetheists should accept both *TIAM-Exclusion* and *TIAM-Exhaustion* for truth-in-a-model so long as they accept their counterparts for truth *simpliciter*, one might wonder what effect it would have on invalidity to reject *TIAM-Exclusion* whilst accepting *Exclusion*. The answer, I think, is nothing: every inference is still invalid, though our strategy for demonstrating this will have to be slightly different, and more akin to the strategy to be employed in the next section against Priest's view. So, in addition to being an unattractive feature of the semantics for independent reasons, it does not help, on its own, with this revenge problem.

On this strategy, we construct special, model-theoretic liar sentences, such as *L* equivalent to:

(*L*) $\neg(M \Vdash L)$

This sentence, informally, says of itself that it is not true in the model *M*. On Priest's view, strictly speaking, *L* will be specified relationally as $\rho(M,\langle L\rangle,1)$, but I stick to the notation just given for simplicity of presentation. If we try to prove a contradiction from this, following similar reasoning to the strengthened liar, we find that the proof breaks down at the point at which the T-scheme would normally be invoked, since we have no equivalent to the T-scheme for truth-in-a-model. What we must do instead is to select *M* carefully, such that something like the T-scheme holds for it. The model we are interested in is the actual model where truth-in-a-model matches exactly truth *simpliciter* in English. It is a great virtue of metatheoretically inconsistent dialetheism allows for the existence of such a model, since otherwise, arguably, truth-in-a-model fails to characterise truth *simpliciter*.

The obvious condition to stipulate in picking out this model, *M*, is that, for any sentence α,

(*) $M \Vdash a \leftrightarrow T\langle a \rangle$

One question arising here is what sort of conditional we should take (*) to contain. I suggest that it ought to contrapose (and so, if '→' is not taken to contrapose, (*) is really $M \Vdash a \leftrightarrow T\langle a \rangle \wedge \neg M \Vdash a \leftrightarrow \neg T\langle a \rangle$). The reason for this is that the conditional must contrapose if truth in *M* is to exactly match (both the positive and negative extensions of) truth simpliciter; if it does not, we are left with the possibility, for any *a*, of having $\neg M \Vdash a$ but failing to have $\neg T\langle a \rangle$ and *vice versa*. In fact, if the conditional does not contrapose, (*) does not specify a unique model: for there are infinitely many models, each satisfying non-contraposable (*), but differing in their assignments of $\neg M \Vdash a$ and $\neg T\langle a \rangle$, respectively.

In any case, I see nothing dialetheically objectionable about the conditional in (*) being contraposable. Certainly there is nothing dialetheically problematic about the mere existence of a model satisfying (*) so-construed. This being so, it seems reasonable that we be able to pick it out as above. It's worth pointing out that Priest's reasons for rejecting the contraposability of the T-scheme do not apply here: one can perfectly well accept the contraposability of (*) and reject the contraposability of the T-scheme.

So there is nothing wrong with us taking (*) in its strongest, contraposable, form. Even if this is rejected, however, this does not avoid the revenge problem. It does, however, collapse the difference, from the point-of-view of this problem, between the view currently under discussion and the view of Priest's discussed in the next section. So, if a dialetheist accepts the transparency of truth *simpliciter*, but rejects *TIAM-Exclusion* as well as the contraposed form of (*), then they will fare exactly as Priest's view does there, though for the reasons given, their view will, additionally be unpleasantly *ad hoc*.

To return to the case where (*) is accepted as contraposable, the proof of contradiction from *L* proceeds as follows:

(1) $M \Vdash L$	(suppose for *reductio*)
(2) $T\langle L \rangle$	(1, (*))
(3) *L*	(2, T-scheme)
(4) $\neg(M \Vdash L)$	(3, substituting equivalents)

(5) $\neg(M \Vdash L)$ (1, 4, *reductio*)

(6) L (5, substituting equivalents)

(7) $T\langle L\rangle$ (6, T-scheme)

(8) $M \Vdash L$ (7, (*))

(9) $M \Vdash L \wedge \neg(M \Vdash L)$ (5, 8, $\wedge$-Intro)

So L both receives 1 in M and fails to receive 1 in M. We can now demonstrate the invalidity of *Reflexivity* by substituting L to obtain the instance $L \vDash L$ in which the premises receive 1 and the conclusion fails to receive 1. In fact, since, on the view under consideration, the T-scheme is transparent, and so *Exclusion* holds, the effect of (*) understood contraposably, is to reinstate *TIAM-Exclusion* in the restricted case of M. So, for any complex sentence, A, built from L using the recursive clauses for the connectives, A both receives 1 and fails to. Moreover, as Priest has pointed out the existence of the trivial model falsifies (at the actual world) every sentence containing a conditional, and so, by the contraposable T-scheme, delivers the untruth of every sentence containing a conditional. By (*), then, every sentence containing a conditional fails to receive 1 in M and so every inference principle whose conclusion contains a conditional is invalid.

So, if the dialetheist accepts that truth is transparent, and accepts *TIAM-Exclusion*, it follows from the existence of the trivial model (or world) that every inference principle is invalid (so long as one accepts Priest's account of the conditional). If, problematically, I have argued, they reject *TIAM-Exclusion* (despite accepting *Exclusion*), then so long as they accept (*) in its contraposable form, it follows from the universal model M and the existence of a liar sentence L, that every inference principle is invalid. If (*) is accepted only in its non-contraposable form, then the situation is the same as is described in the next section.

4.2.2 Invalidity with a Non-Contraposable T-scheme

If the conditional in the T-scheme does not contrapose, the situation is different to that described above. Though it is true that every sentence in the trivial model receives both 1 and 0, in the absence of *TIAM-Exclusion*, we cannot demonstrate that any such sentence additionally fails to receive 1. So the

straightforward strategy of employing the trivial model will not work in this case. Instead, we employ the strategy which invokes the universal model *M*. The proof of a the contradiction $M \Vdash L \wedge \neg(M \Vdash L)$ proceeds exactly as in the previous section, since it does not depend on the contraposability of the T-scheme. So we may use *L* to demonstrate the invalidity of any inference principle which is capable of taking an atomic sentence as its conclusion; not, of course, because *L* is atomic, but because we can obtain an instance of the principle simply by substituting *L* for the single sentence letter in the conclusion. For example, each of ∧-elimination, ∨-elimination (reasoning by cases), *modus ponens*, reflexivity and double negation elimination. The relevant instances being:

(∧-*Elim*) $L \wedge L \vDash L$

(∨-*Elim*) $L \vee L$, $(L \rightarrow L)$, $(L \rightarrow L) \vDash L$

(*MP*) L, $L \rightarrow L \vDash L$

(*Reflexivity*) $L \vDash L$

(*DNE*) $\neg\neg L \vDash L$

In each case, since *L* both receives 1 and fails to receive 1, the premises in each instance receive 1 while the conclusion fails to do so, and each inference principle is invalid.

One might suppose that *L* could be used to invalidate inference principles with more complex conclusions. For example, one might think the following instance would do to eliminate ∧-introduction:

(∧-*Intro*) L, $L \vDash L \wedge L$

Unfortunately, this does not work. The reason is the same as discussed in section (untruth) that the failure of *Exclusion* appears to leave the behaviour of untruth somewhat underdetermined. One would think that, if a sentence, *A*, is untrue that each of $A \wedge A$, $A \vee A$, $\neg\neg A$ would also be untrue (infact, one would think it *obvious* that this is so). In fact, given the failure of *Exclusion*, there is no obvious way of showing that this is the case. I do not think this an advantage to Priest's view. In fact, it seems highly undesirable that untruth is underdetermined in this

way. Though it does make getting revenge on Priest's view slightly more difficult.

What we must do instead, for inference principles with more complex conclusions, is to tailor our liar sentence to the particular inference principle. For example, to invalidate ∧-introduction, we use the sentence *A* equivalent to:

(*A*) ¬(*M* ⊩ *A* ∧ *A*)

Intuitively, *A* says that the result of conjoining it with itself is untrue in *M*. We prove a contradiction from *A* as follows:

(1) *M* ⊩ *A* ∧ *A*	(Suppose for *reductio*)
(2) *T*<*A* ∧ *A*>	(1, (*))
(3) *A* ∧ *A*	(2, T-scheme)
(4) *A*	(3, ∧-elim)
(5) ¬(*M* ⊩ *A* ∧ *A*)	(3, substituting equivalents)
(6) ¬(*M* ⊩ *A* ∧ *A*)	(4, *reductio*)
(7) *A*	(5, substituting equivalents)
(8) *A* ∧ *A*	(6, ∧-introduction)
(9) *T*<*A* ∧ *A*>	(7, T-scheme)
(10)*M* ⊩ *A* ∧ *A*	(8, (*))
(11) (*M* ⊩ *A* ∧ *A*) ∧ ¬(*M* ⊩ *A* ∧ *A*)	(5, 9, ∧-introduction)

So, in *M*, *A* ∧ *A* both receives 1 and fails to receive 1. We can then substitute *A* to obtain the instance of ∧-introduction:

A, *A* ⊩ *A* ∧ *A*

So the premises will receive 1 in *M* and the conclusion will fail to receive 1 and so ∧-introduction is invalid. The case establishing the invalidity of ∨-introduction, *A* ⊨ *A* ∨ *B*, employing the sentence *B*, equivalent to ¬(*M* ⊩ *B* ∨ *B*) mirrors that of ∧-introduction almost exactly. For double-negation-introduction involves the sentence *N*, equivalent to ¬(*M* ⊩ ¬¬*N*), from which a contradiction is proved thus:

(1) *M* ⊩ ¬¬*N*	(suppose for *reductio*)
(2) *T*<¬¬*N*>	(1, (*))

(3) $\neg\neg N$	(2, T-scheme)
(4) N	(3, double-negation-elimination)
(5) $\neg(M \Vdash \neg\neg N)$	(4, substituting equivalents)
(6) $\neg(M \Vdash \neg\neg N)$	(5, *reductio*)
(7) N	(6, substituting equivalents)
(8) $\neg\neg N$	(7, double-negation-introduction)
(9) $T\langle\neg\neg N\rangle$	(8, T-scheme)
(10) $M \Vdash \neg\neg N$	(9, (*))
(11) $(M \Vdash \neg\neg N) \wedge \neg(M \Vdash \neg\neg N)$	(6, 10, $\wedge$-introduction)

Again, we now substitute N into double-negation-introduction to obtain the instance $N \vDash \neg\neg N$ in which the premise relates to 1 and the conclusion fails to and, thus, double-negation-introduction is invalid. N will also serve to invalidate Priest's *reductio* rule, $\alpha \rightarrow \neg\alpha \vDash \neg\alpha$ by substituting $\neg N$ for α.

One question is whether this piecemeal strategy can be developed into a general result. I think, by appeal to a result by Heck (2007)[6],that it can. In his paper, Heck analyses the extent to which genuine self-reference is possible in certain expanded languages for arithmetic. The result in which I am interested is his proof of the *Structural Diagonal Lemma*. In the language PA_s, which is {0, S, +, ×} extended by the addition of a truth-predicate T.

He states the Structural Diagonal Lemma as follows (2007, p.9):

Structural Diagonal Lemma: Let P be a truth-functional schema in (distinct) sentence letters $p_1,...,p_n$. Let $A(x)$ be the substitution instance of P in which each of p_i has been replaced by a corresponding formula $A_i(x)$ containing just x free. Then there is a formula G such that:

(1) G is the substitution instance of P in which each p_i has been replaced by a corresponding formula G_i;
(2) $PA_s \vdash G_i \equiv A_i(\langle G\rangle)$
(3) $PA_s \vdash G \equiv A(\langle G\rangle)$

I refer the reader to Heck's paper (2007, pp.10-12) for the proof, of which Heck says "nothing in the proof actually requires the assumption that $A(x)$ contain

[6] My thanks to Dave Ripley who, in conversation, pointed me in the direction of Heck's paper.

only x free, so [the proof] easily adapts to a proof of the form that allows parameters." (2007, p.10n.11) There is no machinery here which is dialetheically unacceptable and so no reason, that I can see, for a dialetheist to reject the Structural Diagonal Lemma.

This establishes that, for any inference principle whose conclusion contains only extensional connectives, there is a sentence *G*, equivalent to ¬(*M* ⊩ *G*), such that the *G* has the same logical form as the conclusion of the principle. We can then reason in the standard way to the conclusion that *M* ⊩ *G* and ¬(*M* ⊩ *G*), which invalidates the principle. This delivers the result that any inference principle whose conclusion does not contain a conditional is invalid.

This leaves the inference principles which contain in their conclusion a conditional. Since Priest's conditional is not truth-functional, it is not covered by Heck's result. One way one might expect to demonstrate the invalidity of such principles, following the piecemeal strategy above, would be as follows. Consider some inference principle whose conclusion is $P \rightarrow Q$ and take the sentence *B*, equivalent to ¬(*M* ⊩ *B* → *B*). Unfortunately, this does not work, since *B* is not a liar sentence: *B* → *B* is a logically true (only), and so *B*, which says of *B* → *B* that it is untrue, is simply false.

What we might do instead is use the sentence κ, equivalent to ¬(*M* ⊩ *t* → κ), where '*t*' is a truth-constant; most simply, we can simply take '*t*' to be the conjunction of all truths. Two useful features of *t* are that *t* → *a* ⊨ *a*, and *a* ⊨ *t* → *a* both hold, which I will call '*t*-elim' and '*t*-intro', respectively. Given these features, contradiction is proved from κ as follows:

(1) *M* ⊩ *t* → κ	(suppose for reductio)
(2) *T*<*t* → κ>	(1, (*))
(3) *t* → κ	(2, T-scheme)
(4) *κ*	(3, *t*-elim)
(5) ¬(*M* ⊩ *t* → κ)	(4, substituting equivalents)
(6) ¬(*M* ⊩ *t* → κ)	(1, 5, reductio)
(7) κ	(6, substituting equivalents)
(8) *t* → κ	(7, *t*-intro)
(9) *T*<*t* → κ>	(8, T-scheme)

(10) $M \Vdash t \rightarrow \kappa$ (9, (*))

(11) $(M \Vdash t \rightarrow \kappa) \wedge \neg(M \Vdash t \rightarrow \kappa)$ (6, 10, $\wedge$-introduction)

As before, we now have a counter example to any inference principle which can take as its conclusion a conditional with atomic antecedent and consequent. This strategy can be applied in a piecemeal way to a number of inference principles whose conclusions contain conditionals. For a final example, I show how to invalidate the principle $\alpha \rightarrow \beta \vDash (\gamma \rightarrow \alpha) \rightarrow (\gamma \rightarrow \beta)$, which Priest has shown holds for his conditional (it is listed as fact 12 in section 3.2.1). This argument takes advantage of the fact that, since $\vDash t$ and $\vDash t \rightarrow t$, we have $t \vDash t \rightarrow t$ and $t \rightarrow t \vDash t$, so t and $t \rightarrow t$ are equivalent. We employ the sentence γ, equivalent to $\neg(M \Vdash ((t \rightarrow t) \rightarrow (t \rightarrow \gamma))$ and demonstrate that contradiction follows thus:

(1) $M \Vdash ((t \rightarrow t) \rightarrow (t \rightarrow \gamma))$ (suppose for *reductio*)

(2) $T\langle(t \rightarrow t) \rightarrow (t \rightarrow \gamma)\rangle$ (1, (*))

(3) $(t \rightarrow t) \rightarrow (t \rightarrow \gamma)$ (2, T-scheme)

(4) $t \rightarrow (t \rightarrow \gamma)$ (3, substituting equivalents)

(5) $t \rightarrow \gamma$ (4, t-elim)

(6) γ (5, t-elim)

(7) $\neg(M \Vdash ((t \rightarrow t) \rightarrow (t \rightarrow \gamma))$ (6, substituting equivalents)

(8) $\neg(M \Vdash ((t \rightarrow t) \rightarrow (t \rightarrow \gamma))$ (1, 7, *reductio*)

(9) γ (8, substituting equivalents)

(10) $t \rightarrow \gamma$ (9, t-intro)

(11) $t \rightarrow (t \rightarrow \gamma)$ (10, t-intro)

(12) $(t \rightarrow t) \rightarrow (t \rightarrow \gamma)$ (11, substituting equivalents)

(13) $T\langle(t \rightarrow t) \rightarrow (t \rightarrow \gamma)\rangle$ (12, T-scheme)

(14) $M \Vdash ((t \rightarrow t) \rightarrow (t \rightarrow \gamma))$ (13, (*))

(15) $M \Vdash ((t \rightarrow t) \rightarrow (t \rightarrow \gamma)) \wedge \neg(M \Vdash ((t \rightarrow t) \rightarrow (t \rightarrow \gamma)))$ (8, 14, $\wedge$-introduction)

As before, we simply substitute γ into the principle in the obvious way, and so have a counterexample to its validity. It is unclear whether something like this strategy can be extended to a general result for every inference principle with a conditional conclusion; at any rate, I do not have a general result at the moment.

One further possible strategy worth mentioning is one invoking what we might call a ‘super trivial’ model in which every atomic sentence both relates to 1 and fails to relate to 1. In fact, if the dialetheist accepts *TIAM-Exclusion*, this is just the standard trivial model. Priest, however, does not accept this principle, and so in the trivial model he already accepts, every sentence relates both to 1 and to 0, but not all sentences additionally fail to relate to 1. If Priest accepts the existence of such a model (which he may well do), then this immediately invalidates every inference principle which can take an atomic sentence as its conclusion. However, it is not clear that it invalidates anything else, since (as already discussed) the failure of *TIAM-Exclusion* would seem to block the obvious arguments for the untruth of complex sentences from the untruth of their component sentences. So even though dialetheists may allow for the existence of such a model, and it will certainly invalidate some inference principles, the more complicated strategy outlined above, which invokes the model *M*, may still be required to spread inconsistency to more complex principles.

So, on Priest’s strategy, both *Exclusion* and *TIAM-Exclusion* fail. This rules-out the strategy from the previous section (5.2.1) which employed the trivial model. Instead, we used the strategy which employs the universal model *M* and constructs ‘bespoke’ liar sentences for each inference principle we wish to invalidate. This allowed us to invalidate every inference principle whose conclusion can be an atomic sentence, a conjunction of atomic sentences, a disjunction of atomic sentences or a negated atomic sentence (as well as double-negation introduction). This was extended to every inference principle with a conditional-free (and so extensional) conclusion by a result due to Heck (2007). Finally, any principle with a conditional conclusion, able to take atomic sentences as antecedent and consequent, is also invalid. This can be extended in a piecemeal fashion to some principles with more complex conditional conclusions, but, as yet, no more general result has been demonstrated.

Whether, in the end, Priest’s view, by rejecting *Exclusion* and *TIAM-Exclusion*, fares any better here than versions of dialetheism accepting those principles depends on whether the stronger result, that every inference principle whose conclusion contains a conditional is invalid, can be established. It seems to me that this result should be obtainable, but even if it is not, Priest’s view is not *much* better off. It is still the case that almost every inference principle is

invalid, including all the introduction and elimination rules for the logical constants, and all of the principles we are likely to find in a logic textbook, or use in everyday reasoning.

4.2.3 Revenge and the Inexpressibility of Invalidity

The fact that validity is dialetheically inconsistent (at least for versions of dialetheism with a paraconsistent metatheory) is interesting and important, all the more so because of the extent of its inconsistency. At worst, every inference principle is invalid, and at best, almost every inference principle one is likely to encounter is invalid. The revenge problem this generates is that is renders dialetheists unable to express invalidity. One of the crucial features of dialetheism is that it is, we think, non-trivial: dialetheists are able to accept inconsistency, but also contain inconsistency. So it is a crucial semantic fact about dialetheism that principles such as explosion and disjunctive syllogism, which would otherwise trivialise the theory, are invalid. If dialetheists cannot express this fact, then their view suffers from expressive limitations of the very kind they argue that non-dialetheists fact: that is, they have a revenge problem.

Consider, first, the case in which every inference principle is invalid. This is the position of metatheoretically inconsistent dialetheism in which *TIAM-Exclusion* holds, and perhaps also Priest's view, pending some more definitive result concerning principles whose conclusions contain conditionals. In this case, 'invalid' and 'inference principle' are logically equivalent notions. So, for all the distinction the notion draws when they say 'Disjunctive syllogism is invalid', they may as well have said 'Disjunctive syllogism is an inference principle'. Invalidity, here, is completely vacuous, and so does not express what is intended about disjunctive syllogism; the feature predicated of disjunctive syllogism is possessed by *every* inference principle trivially.

Just as, with the introduction of dialetheism about truth, it became necessary to introduce notions like 'just true' and 'just false' to characterise the sentences which are not dialetheia, it seems reasonable to want similar notions once we recognise that validity is dialetheic. So we may wish to introduce notions such as 'just valid' and 'just invalid'. In fact, the former notion is redundant, in this

case, since no inference principle is just valid. Still, though every principle is invalid, what is special and bad about principles like disjunctive syllogism and explosion is that they are *just* invalid, whereas the legitimate principles, like reflexivity and *modus ponens*, are both valid and invalid.

The obvious definition of 'just invalid', of course, is 'invalid and not valid'. Unfortunately, since 'invalid' and 'not valid' are equivalent notions, 'invalid' and 'just invalid' are also equivalent. So, on this obvious definition, 'inference principle', 'invalid' and 'just invalid' are logically equivalent. So, on this version of dialetheism, both invalidity and just-invalidity are vacuous notions which fail to pick out the feature of disjunctive syllogism that we are interested in.

So, if dialetheists accept *TIAM-Exclusion*, they have a revenge problem. Specifically, they cannot express the crucial semantic facts about their theory which require for their expressibility non-vacuous notions of invalidity and just-invalidity.

Assuming that we cannot establish the stronger result mentioned above, Priest's view (or, more generally, views in which *TIAM-Exclusion* fails) is different, but only slightly. It is not the case on theories of this kind that 'invalid' and 'inference principle' are logically equivalent. The overlap, however, is still enormous. The notion being employed in sentences such as 'Disjunctive syllogism is invalid' still applies to every inference principle whose conclusion lacks a conditional, and a number of others, including reflexivity, *modus ponens*, $\wedge$-introduction, and almost every other inference principle anyone ever uses. It is still the case, on this version of dialetheism, that 'just invalid' is equivalent to 'invalid' and so each of these principles is also *just* invalid.

So, even if dialetheists reject *TIAM-Exclusion*, and assuming the stronger result cannot be established for these theories, they still have a revenge problem. Both 'invalid' and 'just invalid' are equivalent, so these theories cannot express the latter as distinct from the former. Moreover, the overlap between these notions and validity is so significant that they are deprived of their usefulness and so cannot be used to express the crucial semantic facts about dialetheism which require either consistent, or at least less inconsistent, notions of invalidity.

4.2.3.1 Invalidity Revenge and Just False Revenge

It is worth briefly discussing the differences between the 'just false' problem already extant in the literature and the invalidity revenge problem I have just introduced, in case it be thought that nothing new has been added.

The first difference is that the 'just false' problem concerns only the notion 'just false', leaving 'false' alone, and not threatening any further, unpalatable properties of 'false' that the dialetheist has not already happily accepted. The invalidity revenge problem, on the other hand, concerns both invalidity and just-invalidity, demonstrating that the notions are both equivalent and vacuous (or near enough).

This being the case, one might think the invalidity revenge problem concerns a more central notion than does the just false problem. The notion 'just false' only entered the debate surrounding truth and paradox when dialetheism became an important participant and the central notions of truth and falsity became inconsistent. But it is not unreasonable to characterise logic as being the study of (principally deductive) reasoning and how we ought to distinguish correct (valid) from incorrect (invalid) reasoning. So the notion of invalidity is one of the defining notions of logic, and this is one of the notions which the invalidity revenge problem threatens to demonstrate dialetheically inexpressible.

A further difference is that the problematic overlap in the case of the invalidity revenge problem is far greater than in the case of the just false problem. If the dialetheist accepts *Exclusion*, then the obvious definition of 'just false' collapses the notion into 'false', but this does not extend the overlap between these notions and truth. With the invalidity revenge problem, on the other hand, 'invalid' and 'just invalid', ordinarily defined, are equivalent, and the overlap between these notions and validity extends to cover all valid inference principles, on the assumption of *TIAM-Exclusion*: 'invalid', 'just invalid' and 'inference principle' are logically equivalent. In Priest's case, his rejection of *Exclusion* allows for falsity and just-falsity to be inequivalent, and for a significant number of standard self-referential, paradoxical sentences to be true and false, but to fail to be just false. But with the invalidity revenge problem,

rejecting *TIAM-Exclusion* may not help at all, and if it does, it helps very little. Invalidity and just-invalidity are still equivalent and the notions apply to either every inference principle or, at best, almost every inference principle one is likely to encounter.

4.3 Avenues of Response

I now turn to how a dialetheist might respond to the invalidity revenge problem I have characterised. One possible response would be to accept that invalidity is vacuous and to be given up, but to define 'just invalid' in some new way to express what is wrong with inferences like disjunctive syllogism. This will, I think, be parasitic on some more general, consistent redefinition of 'just false' (or, perhaps 'just untrue'). The reason for this is that every counterexample to a valid inference principle described in the previous section consisted in finding some sentence which, in some model, was both true and untrue and inserting it into the inference principle in question such that the premises were true, but the conclusion untrue. In each case, because the inferences were valid too, the conclusion, in addition to being untrue, was also true (again, in the model). So the only time we have counterexamples to valid inference principles is when we construct cases in which the premises are true while the conclusion is both true and untrue. This is different from inference principles which are *just* invalid, because in those cases we have counter-instances in which the premises are true and the conclusion false only (or untrue only).

So what the dialetheist wants here is some notion, some form of negation, which expresses that a sentence fails to be true, where this must be understood consistently (i.e. it cannot be such that something is untrue in this sense, but also true). We call this new notion *NEG* and define an inference to be invalid, $A \nVdash B$ just if there is some model I such that $I \Vdash A \wedge NEG\text{-}(I \Vdash B)$. We can also define a sentence, α, to be just false if and only if $F\langle\alpha\rangle \wedge NEG\text{-}T\langle\alpha\rangle$. But, of course, we can also construct a sentence, β, equivalent to:

(β) $NEG\text{-}T\langle\beta\rangle$

From which we require very little to deduce $T\langle\beta\rangle \wedge NEG\text{-}T\langle\beta\rangle$, defeating the purpose of introducing *NEG* in the first place. An argument for this from β is as follows;

(1) $T\langle\beta\rangle$	(suppose for *reductio*)
(2) β	(1, T-scheme)
(3) $NEG\text{-}T\langle\beta\rangle$	(2, substituting equivalents)
(4) $NEG\text{-}T\langle\beta\rangle$	(3, *NEG-reductio*)
(5) β	(4, substituting equivalents)
(6) $T\langle\beta\rangle$	(5, T-scheme)
(7) $T\langle\beta\rangle \wedge NEG\text{-}T\langle\beta\rangle$	(4, 6, $\wedge$-introduction)

All that is required for this argument, beyond principles already accepted by dialetheists is that a *reductio* rule is valid for *NEG*. This need only be the weak reductio rule Priest already accepts for negation, $\alpha \rightarrow NEG\text{-}\alpha \vDash NEG\text{-}\alpha$. It is difficult to see how the failure of such a rule could be motivated (at least, in the present context), but there are other arguments for $T\langle\beta\rangle \wedge NEG\text{-}T\langle\beta\rangle$ which don't involve it anyway. For example:

(1) $T\langle\beta\rangle \vee NEG\text{-}T\langle\beta\rangle$	(*NEG-LEM*)
Case 1:	
(1a) $T\langle\beta\rangle$	
(1b) β	(1a, T-scheme)
(1c) $NEG\text{-}T\langle\beta\rangle$	(1b, substituting equivalents)
(1d) $T\langle\beta\rangle \wedge NEG\text{-}T\langle\beta\rangle$	(1a, 1c, $\wedge$ -introduction)
Case 2:	
(2a) $NEG\text{-}T\langle\beta\rangle$	
(2b) β	(2a, substituting equivalents)
(2c) $T\langle\beta\rangle$	(2b, T-scheme)
(2d) $T\langle\beta\rangle \wedge NEG\text{-}T\langle\beta\rangle$	(2a, 2c, $\wedge$-introduction)
(2) $T\langle\beta\rangle \wedge NEG\text{-}T\langle\beta\rangle$	(1, 1d, 2d, reasoning by cases)

This argument, too, might be blocked, perhaps by denying that *NEG* obeys the law of excluded middle, or that reasoning by cases is valid, but again, there

doesn't seem much motivation for this, and the more principles of this kind which fail for *NEG*, the less it looks like a form of negation at all.

Perhaps there is some way to characterise a consistent form of negation for dialetheists whose metatheory is paraconsistent, but it is not at all obvious how this would be achieved, and it has not, to my knowledge, been seriously attempted thus far.

One attempt which has been made for dialetheists with a consistent metatheory is worth discussing for illustrative purposes. Franz Berto, in, for example, his (2014) has developed a primitive notion of metaphysical exclusion, characterised as a predicative functor, holding between properties which metaphysically exclude one another. Though the notion is primitive, and so has no explicit definition, there are still formal principles which hold of it. One such is the principle in the metatheory that a property excludes another if and only if their extensions are (necessarily) disjoint. Since Berto's metatheory is consistent, this principle does not make exclusion inconsistent in the object theory. It is difficult to see how anything like this principle could hold in the present, metatheoretically inconsistent context (where extensions are allowed to be both disjoint and not) without thereby introducing inconsistency into the notion of exclusion. It is not required, of course, that a *definition* be given of a notion which is supposed to be primitive, but we should still expect some formal constraints to hold of exclusion, and it is unclear what these could be such that the notion doesn't turn out inconsistent.

Still, supposing some constraints could be given to make the notion sufficiently wieldy, we might then hope to characterise an invalid inference, $A \nvDash B$ as being such that there is some model I such that A is true in I and A's truth in I excludes B's truth in I.

This might work for some invalid inference principles, for example $A \vDash B$; presumably, so long as the truth of some sentences excludes the truth of some others, there is some instance of this such that the truth of the premise excludes the truth of the conclusion. The notion is less helpful in other, more interesting cases, however. The reason is that the obvious counter-examples to the just-invalid inference principles in which we are interested, for example, disjunctive

syllogism or explosion, involve necessarily dialetheic sentences arising from semantic paradoxes like the liar. For example, a counter-instance to disjunctive syllogism might be $\neg L$, $L \vee A \vDash A$, where L is some liar sentence and A is some sole-falsehood, such as 'Gareth Young is 10-feet-tall'. But since the liar sentence L is, presumably, necessarily true and false, it is difficult to see how it could metaphysically exclude anything, and, in particular, how it could be thought of as excluding my being 10-feet-tall. All the paradigmatic dialetheia are of this necessary sort, so it does not appear that metaphysical exclusion could help with this problem, even if a satisfactory account of the notion could be made in the metatheoretically paraconsistent context.

Another strategy dialetheists might wish to pursue in response to the invalidity revenge problem appeals to triviality. One problematic feature of inference principles like disjunctive syllogism and explosion, and the main reason we want to be able to express their invalidity, is that, were they valid, the theory would trivialise. So it seems reasonable to wish to invoke this fact in our response to the problem. The response would be to give up on notions like invalidity and just-invalidity, but to express that a certain inference principle is bad by pointing to the fact that, were it added to the theory, the theory would entail everything.

One might think that, rather than answering the objection, this simply underlines it. It is of course correct that these principles would be trivial were they valid, but this is the very reason we want to be able to say they are *invalid*. If I notice, when first introduced to dialetheism, that there is a straightforward argument from dialetheism to triviality via disjunctive syllogism, and inquire as to the principle's validity, it will hardly assuage my concerns if the dialetheist simply replies that the principle does indeed entail triviality.

In any case, this response will not work, since not all invalid inferences would trivialise if added to dialetheism. To take Priest's theory as an example, not all inferences he rejects are rejected because they would, if he accepted them, entail everything. For example, for all α, $T\langle\neg\alpha\rangle \vDash \neg T\langle\alpha\rangle$, which is a rule form of what Priest calls 'Exclusion', is rejected by appeal to his teleological account of truth. It is, for Priest, invalid, but it would not, so far as I can see, trivialise his theory if he accepted it. For another example, related to the previous, Priest

rejects the contraposability of the conditional: $P \rightarrow Q \vDash \neg Q \rightarrow \neg P$ is, for him, invalid. But the reasons for this, again, are not the threat of triviality, but philosophical concerns about what conditional ought to appear in the T-scheme. For a final example, one obviously bad inference principle is $A \vDash$ Priest is a fried egg. This would, if valid, allow us to conclude that Priest is a fried egg from any sentence, but it would not, so far as I can tell, entail *everything*. It might be objected that since the conclusion of the principle is an interpreted sentence that it is not an inference principle, but this doesn't matter. It is still a putative principle of some kind, by which one might reason, and which is bad, because it is invalid. This fact ought to be able to be expressed dialetheically, but cannot be if our only resources for expressing what's bad about invalid inferences is to point to their triviality.

4.3.1 Invalidity Revenge on the Model-Theoretic Strategy

The characterisation of the invalidity revenge problem outlined above has assumed that the dialetheist's metatheory is carried out along the lines of the relevant strategy, which I have suggested is the most promising option for the dialetheist. Priest thinks another, model-theoretic strategy is also available. I have argued that this strategy will not work. Still, it is worth briefly discussing the extent to which the invalidity revenge problem affects this strategy.

As discussed, the only conditional available on this strategy is material (though whether, given the failure of *modus ponens*, this ought to be called the material 'conditional', is contentious). Since this does not detach, and it features in the definition of validity, validity itself does not detach (which is to say just that we cannot 'detach' the truth of the conclusion of a valid inference from the truth of its premises). Priest's response to this difficulty is to appeal to the fact that classical reasoning can be recaptured in consistent contexts. So, unless the premises of an inference are both true in a model and untrue in the same model, we are at liberty to detach the conclusion. But if the premises are both true and untrue in the same model, we cannot detach the conclusion. Because the invalidity revenge problem essentially involves sentences which are both true and untrue in the same model, one might think this makes things more difficult, since reasoning from such contradictions is almost impossible. Indeed, one might

wonder how we even demonstrate the existence of such contradictions if our ability to reason about them is crippled in this way.

For example, consider the simplest model-theoretic liar sentence described above, *L*, equivalent to ¬(*M* ⊩ *L*). As before, we derive a contradiction as follows:

(1) *M* ⊩ *L*	(suppose for *reductio*)
(2) *T*<*L*>	(1, (*))
(3) *L*	(2, T-scheme)
(4) ¬(*M* ⊩ *L*)	(3, substituting equivalents)
(5) ¬(*M* ⊩ *L*)	(1, 4, *reductio*)
(6) *L*	(5, substituting equivalents)
(7) *T*<*L*>	(6, T-scheme)
(8) *M* ⊩ *L*	(7, (*))
(9) *M* ⊩ *L* ∧ ¬(*M* ⊩ *L*)	(5, 8, ∧-introduction)

For each step in this argument, if we want the truth of the conclusion, we need to be able to obtain the truth of the conclusion of a valid argument from the truth of its premises but, if *L* is contradictory as in (9), we can't do this. Still, we may think of this as providing a *reductio* of the supposition that there are no sentences which are both true and untrue in the same model, since, if there were no such sentences, could *always* obtain the truth of the conclusion of a valid argument from the truth of its premises, legitimating each step of the argument just given. So, if we suppose that there are no such sentences, we can demonstrate that there are some. In fact, the argument above seems to demonstrate that *L* itself must be of this kind since, if it were not, each step in the argument above would be licit and we could obtain the truth of (9). The only way for the argument to fail is for the conclusion, that *L* is both true and untrue in *M*, to be correct after all. The same will be true of each of the similar arguments given in the preceding section. So, in fact, there exist a great many sentences which are both true and untrue in the same model, and all the ones we used to invalidate dialetheically valid inference principles are among them.

What we need to get from these sorts of sentences to the invalidity of a given inference principle, assuming that argument above gives us only {*M* ⊩ *L*, ¬(*M* ⊩

L)} is an instance of ∧-introduction, to give us $M \Vdash L \wedge \neg(M \Vdash L)$, then an instance of the principle, $A \wedge \neg B \vdash \neg(A \rightarrow B)$, giving us $\neg(M \Vdash L \rightarrow (M \Vdash L))$. This would, for instance, be a counter-instance to reflexivity. One might think that the fact that the conditional in the definition is material, and that the premises of this instance are both true and untrue in a model, would mean that these steps are illicit. This is not so. The definition of validity is such that, in certain cases, we cannot detach the truth of the conclusion of a valid argument from the truth of its premises. But this does not prevent us applying syntactic inference rules like ∧-introduction, to obtain the counter-example we want.

Pending some argument that the weakness of the materially-defined notion of validity prevents, not just obtaining the truth of the conclusion of a valid argument from the truth of the premises, but the application of syntactic inference rules, we can demonstrate the invalidity of some valid inference principles, *even on the model-theoretic strategy*. Indeed, it would seem, all the same ones as before. Given the weakness of this strategy, this is a somewhat surprising result.

4.4 Chapter Conclusion

The purpose of this chapter has been argue that versions of dialetheism which take their metatheory to be paraconsistent suffer from revenge problems. Priest's dialetheism is the most sophisticated and well-described version of dialetheism in general, and this sort of dialetheism, more specifically. Priest gives powerful arguments for this view, which were critically discussed in Chapter 3. Chief among this was the problem of revenge: each of the extant theories of the paradoxes was argued by Priest to face self-refuting expressive limitations and dialetheism, he claimed, is the only way to avoid these. Other arguments, too, were described, for example that naive set-theory, paraconsistently construed, is the only account of sets able to respect our mathematical practices. My discussion of revenge problems for the view first focused on the alleged inexpressibility of the notions 'non-dialetheia' and 'just false'. After giving a couple of possible responses to these problems, I moved on to my own 'Invalidity Revenge' problem, showing that, for dialetheist with a paraconsistent metatheory, either all, or almost all inference principles are

invalid and, moreover, that ‘just-invalid’ is equivalent to ‘invalid’. This renders these notions inexpressible and, I have argued, means that dialetheism has a revenge problem of the same kind they charge non-dialetheists as facing. If this is correct, it is a significant blow to the motivations for dialetheism. Many philosophers, I suspect, would think the demonstration of a revenge problem for dialetheism to, effectively, refute the theory. This is not my view. Though, of course, the revenge problem is extremely serious, plausibly, every available view of the paradoxes suffers from it. It may be that dialetheism, involving, as it does, the acceptance of true contradictions, requires some special theoretical benefit to outweigh this special cost. Revenge immunity would certainly be such a benefit. However, it may not be the only one available. In particular, the set-theoretic arguments of Priest’s, and especially those concerning absolute generality, seem to me quite powerful. Though these are arguments for naïve set theory, rather than dialetheism *per se*, treating set theory inconsistently is certainly a strong contender for a naïve account of the sets. A proper assessment of these arguments, and whether dialetheism can really be motivated by them, is beyond the scope of this thesis. However, if dialetheism, of the kind endorsed by Priest, has a revenge problem, these arguments seem the most promising as an alternative motivation for the view.

Chapter 5: Metatheoretically Consistent Dialetheism

5.0 Introduction

This chapter is about dialetheism which takes as its metatheory classical *ZF* set theory. The primary focus will be the version of this view defended by JC Beall in his 2009 book, *Spandrels of Truth*. A number of authors have defended (or at least sympathetically characterised) variants of dialetheism in a number of areas, for example, Cogburn (2004), Garfield (2004), Kroon (2004), Mares (2004), Mortensen (in a number of works including his (1985), (1997), (2009)) and Ripley (2011). They have not, however, offered a comprehensive theory of dialetheism, specifically endorsing a particular metatheory and responding to potential revenge objections. Since Beall's book is the only one I am aware of to do this explicitly, with a classical metatheory, it seems right to focus on this.

It is worth mentioning two approaches to dialetheism which won't receive significant discussion here. The first is dialetheism underpinned with a logic in which some structural meta-rule fails (these are known as sub-structural logics), such as cut, contraction or transitivity. A version of sub-structural dialetheism is defended, for example, by Ripley in his (2013) in which he rejects the transitivity of entailment. Another view which won't be discussed is the co-called detachment-free dialetheism, in which *modus ponens* is rejected as invalid, defended recently by Beall (2013).

One reason these I don't discuss these is lack of space: one can't discuss everything and the views to which I devote most discussion, Priest's and Beall's (at least, the version defended in his 2009), are, at the moment, far more prominent in the literature. Secondly, these views are very new and their authors have not yet offered responses to the revenge problem on behalf of them. Relatedly, since these views have only recently developed, it is not, to my knowledge, known whether they can be given a metatheory which has the same logic as the object theory. This being the case, we should take the metatheory, for the present, to be classical. This fact, I think, means that these views should be subject to at least some of the revenge problems which I argue face

metatheoretically consistent dialetheism of a kind more similar to Beall's (2009). It seems to me, for example, that the formal revenge problem outlined below will apply to these theories in much the same way as to the views I discuss explicitly below. If Ripley's theory can be given an adequate metatheory in sub-structural terms, things may be different, but we shall have to wait for this development before taking a view. Similarly, Beall may be able to give a detachment-free metatheory for his detachment-free object theory. Though this might depend on the details of the theory, we should note the similarity between this project and what Priest calls 'The Material Strategy' for a dialetheist metatheory (discussed in section 4.1 of Chapter 4), which also did not contain a detachable conditional, to its disadvantage. It's difficult to see how Beall's detachment-free theory would fare any better on this score, so perhaps there is reason to be sceptical of a revenge-free theory along detachment-free lines.

I begin by giving Beall's view in informal terms, giving his account of dialetheia as 'spandrels of truth' which he takes to be inevitable by-products of the logical role which constitutes the truth predicate's sole purpose. I then present the formal semantics Beall gives for his view, including what he takes to be the correct conditional. Along the way, I offer some critical discussion of Beall's version of metatheoretically consistent dialetheism and of the view more generally.

5.1 Spandrels of Truth

According to Beall, truth is a logical device constructed to overcome our expressive limitations as finite beings. This logical device gives rise to the truth of some contradictions but only, he thinks, in "a fairly mundane, 'deflated' sense." (2009, p.1)

God, because of his infinite capacities, says Beall, could use the truth-free fragment of English to completely describe our world. But because we are unlike God in respect of our capacities, we cannot do this. We must introduce a device which allows us to express generalisations we could not otherwise express. The

device in question is a transparent truth predicate, which is introduced via rules of substitution: *Tr*<α> and α are intersubstitutable in all non-opaque contexts.

According to Beall's picturesque story, before the truth predicate was introduced, we spoke only in a truth-free language. We could express a good number of things (taking his examples), that Max is a cat, say, or that Gödel and Tarski were independently ingenious and so on. But, because of our interests, especially, but not only, our theoretical interests, we must also be able to generalise in a way which is not possible for us without our logical generalising device. I may wish to agree with everything Bob says, for example. If he says only a couple of things, and I know what they are, I can do this by simply repeating them. But if Bob says a great many things, or I haven't heard everything Bob has said (and I wish to agree with him on trust), for example, this may become difficult or impossible. The difficulty is even clearer in theoretical cases: I may wish to agree with quantum mechanics, on the basis of the testimony of experts, despite not understanding the theory, or knowing all of its contents. I could not do this by repeating each of the sentences which make up quantum mechanics. A further problem is that many complex theories might contain infinitely many sentences, in which case it would be *impossible* for me to endorse them by uttering each of such sentences. Instead, Beall thinks, we introduce our transparent predicate, via its substitution rules, and say 'Quantum mechanics *is true*', or 'Everything Bob says *is true*'.

So, although God could agree with a theory containing infinitely many sentences by simply repeating every sentence in the theory, perhaps because he has an infinite amount of time to do it in, or perhaps because he can say each sentence twice as quickly as the last, we cannot. So we introduce a truth predicate to do our generalising for us. This, according to Beall, is the only job truth has in our language: it is not a substantive property which picks out some important feature about the world, but rather a logical tool for the purposes of generalisation.

To illustrate the point, he says we could have achieved the same goal via the story of Aiehtela (pronounced, he says, 'eye-ah-tell-ah') and Aiehtelanu (pronounced 'eye-ah-tell-ah-noo'), where 'Aiehtela accepts *x*' is everywhere (apart from opaque contexts) intersubstitutable with *x* and 'Aiehtelanu accepts

x' is equivalent to the negation of *x*. Whether we use '*x* is true' or 'Aiehtela accepts *x*' doesn't matter, according to Beall, so long as they have the correct, transparent properties; what matters is the logical, generalising job these predicates allow us to do.

When we first learned of Aiehtela and Aiehtelanu, it would, according to Beall, have been natural to reify the characters and ask after their nature. Perhaps we might suppose Aiehtela accepts sentences on the basis of their coherence with other sentences, for example. But, Beall thinks, if we are right to think that the characters were invoked merely for their usefulness in allowing us to generalise, this reification is a mistake: Aiehtela has no essential nature, at least not in the sense that it picks out an important feature of the world.

The moral of these stories given by Beall is, for the most part, the familiar view known variously as 'deflationism', as Beall points out, citing Field's defence (1994) as one with which he is in broad, methodological agreement.

Throughout his (2009), Beall distinguishes his minimal, deflated notion of truth from more substantial notions by calling it 'ttruth' (for 'transparent truth'). To avoid multiplying terminology, I stick to the more usual 'truth', allowing context to make it clear when I am talking about Beall's transparent notion and when more substantial notions are being considered. He doesn't give any extended defence of the view against the well-known problems it faces, since his purpose is simply to give what he thinks is the best account of the paradoxes available to someone who holds this view of truth. Since my purpose is just to assess the extent to which his view (and other metatheoretically consistent forms of dialetheism) faces revenge problems, I follow him in not spending time on objections to deflationism. I am fairly sympathetic to the basic view anyway.

One point worth making is that one might get the impression, from Beall's claim that we could have done without a truth predicate, if we had been less ambitious in our desires to generalise, that we would thereby have done without contradiction. In other words, one might think, for Beall, that our dialetheism is contingent upon us actually adding a truth predicate to our language. That this is so is far from clear. Beall is committed, as are most dialetheists, to what I'll call the 'modal law of non-contradiction):

(*MLNC*) $\neg\Diamond(P \wedge \neg P)$

The effect of this is, so long as a contradiction is merely *possible*, it generates a further, *actual*, contradiction, viz. that the contradiction in question is both possible and impossible. For example, suppose that there is some α for which $\Diamond(\alpha \wedge \neg\alpha)$. By *MLNC* (and the introduction rule for conjunction), this gives us $\Diamond(\alpha \wedge \neg\alpha) \wedge \neg\Diamond(\alpha \wedge \neg\alpha)$. So, for dialetheists, merely possible contradiction generates actual contradiction (this has been pointed out by Slater (1995, pp.451-454) and Restall (1997, pp.156-163)). So, even if we had not introduced our truth predicate, presumably we *could* have, and so there *could* be a sentence, α, such that α and $\neg\alpha$. Since, by *MLNC*, there also *could not* have been such a sentence, we would seem to have actual contradiction whether we introduce the truth predicate or not. So, in fact, dialetheism is more deeply built-in to Beall's view than one might suppose, from the remarks that we could, had our interests been different, done without the truth predicate, which Beall takes to be the source of dialetheia.

The point that possible contradictions generate actual ones is very important for Beall's view, and I will return to it.

Two important logical principles are the law of excluded middle (LEM) and bivalence (BIV):

(LEM) $\vDash \alpha \vee \neg\alpha$

(BIV) $\vDash T\langle\alpha\rangle \vee T\langle\neg\alpha\rangle$

Beall uses '$\vdash$' to record validity, but to keep my usage of symbolism consistent, I use '$\vDash$' for this, keeping '$\vdash$' for syntactic consequence. Assuming the equivalence of falsity with untruth (accepted by all participants in this debate, including Beall), BIV is equivalent to the, perhaps more familiar, form of bivalence $\vDash T\langle\alpha\rangle \vee F\langle\alpha\rangle$. Moreover, assuming truth behaves transparently, LEM and BIV are equivalent too. Beall accepts this and so accepts both. This means his account of negation is *exhaustive* (though, as will be clear when I present the semantics for his view, it fails to be exhaustive at some *abnormal* worlds).

Beall does not give any extensive arguments for this feature of negation against, for example, Field's (2008), which achieves transparent truth without this. Beall admits this and says he does not know of any powerful arguments for preferring his own approach over Field's. His own reason for preferring his view, he says, is a feeling that Field misses an important feature of negation (presumably exhaustion, though one can imagine Field responding that Beall misses an even more important feature, *exclusion*, when he allows *a* and ¬*a*), and that Field's view is too complicated (2009, p.x). However, he says, he does not know how to develop these into arguments against Field (he does, however, devote all of Chapter 4 of his (2009) to discussion of such views). So, for the purposes of discussion, I simply grant, for the present, that LEM (and, equivalently BIV) holds.

Beall employs the term 'spandrels' to describe the dialetheia (though he prefers the term 'glut') which he thinks arise from the semantic paradoxes. The term comes from architecture and is also used in evolutionary biology, and denotes an unintended, but inevitable, by-product of some intended thing. In architecture, archways, as curved figures, inevitably generate more-or-less triangular shapes at the corners of the rectangles in which they are set. These inevitable shapes are spandrels: unintended, but inevitable by-products of archways. Evolutionary spandrels are features of an organism which are not directly selected for, but which unavoidably come along with some feature which is selected for. Beall's example is the male nipple, which is a spandrel of the female nipple.

Dialetheia, Beall thinks, are spandrels of language. They are an unintended, but inevitable, by-product of our transparent truth predicate. We need to be able to make the sorts of generalisations described above, so we introduce our predicate, $T\langle\alpha\rangle$, as intersubstitutable with α, into our language (in this case, English) and, along with the other features of that language, we end up with sentences like *L*:

L: The sentence *L* is not true

Since truth is intersubstitutable, it supports the rules he calls 'Capture' and 'Release':

Capture: $T\langle\alpha\rangle \vDash \alpha$

Release: $\alpha \vDash T\langle\alpha\rangle$

These rules, delivered by transparent truth (as well as the, presumably innocuous, assumption that $\alpha \vDash \alpha$, as well as the standard structural rules), along with the assumption we've made that negation is exhaustive, deliver the existence of dialetheia[7] i.e. they entail that the sentence L is such that $T\langle L\rangle$ and $T\langle\neg L\rangle$.

So Beall is a dialetheist, but thinks of himself as such only in a mundane, deflated sense. One, perhaps small, issue arising here concerns Beall's characterisation of his dialetheism as "mundane" and "deflated" (expressions he has used together at least twice in his book, once on p.1 and once on p.6). He says later that "[t]he position is dialetheic, but only in a modest, 'deflated' way; the '[true] falsehoods' are merely transparently true and transparently false; they are [true] sentences whose negations are also [true]." (2009, p.17)

This might give the impression that his deflationism about truth – that truth is some non-substantial logical tool, rather than a full-blooded property (in some sense to be specified) – has somehow 'deflated' his dialetheism and made this non-substantial. But it's not clear that deflationary dialetheism, *per se*, is any less full-blooded a version of the view than one which takes truth to be a more substantial property. Both think that things of the form $T\langle\alpha\rangle \wedge F\langle\alpha\rangle$ hold, and given the definition of falsity, $T\langle\alpha\rangle \wedge T\langle\neg\alpha\rangle$ would then hold, and given Beall's views about the truth predicate's transparency, these are equivalent to $T\langle\alpha\rangle \wedge \neg T\langle\alpha\rangle$, and to $\alpha \wedge \neg\alpha$. We should note, as an aside, that, for example, a correspondence theorist is perfectly entitled to these equivalencies too. But why should *this* make Beall's *dialetheism* any less robust, or radical, than that of any other dialetheist? Dialetheism is just the view that some contradictions are true, and Beall subscribes to this every bit as much as someone like Priest, who does not endorse deflationism about truth.

It is true, as well shall see, that Beall's dialetheism is less radical, or at least less thoroughgoing, in some ways, than other versions of the view (for example, Priest's). But this is so because of the restricted domain in which he thinks dialetheia arise (an issue on which people accepting his minimalist account of

[7] This is not to say that each of these assumptions are *required* to demonstrate the existence of dialetheia, since we can prove a contradiction from the liar without the use of the law of excluded middle.

truth may differ), not because deflationism about truth somehow, on its own, makes the claim that contradictions are true any less substantive or radical. It may be that Beall does not intend to suggest this further claim, but the possibility of reading him this way seems plausible enough to point out the fault in it.

The restricted domain in which dialetheia arise is, for Beall, that which contains semantic notions like truth. For him, then, dialetheia are a 'merely semantic' phenomenon, and his dialetheism is 'merely semantic' dialetheism. Making more precise this characterisation of his restricted version of dialetheism is the final part of his philosophical view which I will characterise, but I will follow Beall in giving a brief, formal interlude to specify some of the most basic features of the semantics he endorses.

5.2 Basic Formal Picture

Beall calls disjunction, conjunction and negation the 'Boolean connectives'. He does not mean us to understand him as committing himself to the classical behaviour of these connectives, but he wishes to avoid the more usual term 'extensional connectives' since negation, on his view, is not extensional.

He gives the semantics for the Boolean connectives in terms of worlds, invoking the so-called 'Routley star', following the Routleys (1972). I follow Beall's presentation (2009, pp.7-) closely in what follows, and so begin by giving the classical semantics for the Boolean semantics where classicality is ensured by adding a constraint which, when relaxed, gives us the more general view which Beall endorses.

Interpretations are taken to be ordered quintuples $<W, N, @, *, \Vdash>$, where W is the (non-empty) set of worlds, N is the (non-empty) set of normal worlds, and is a subset of W, '@' denotes the actual world and we take the difference of W and N to be the set of abnormal worlds (which may or may not be empty). The star operator '*' on W is such that $w^{**} = w$. Beall uses '$\models$' to indicate that a sentence is true at a world, but to keep notation consistent with other chapters, I use '$\Vdash$' instead, and so '$w \Vdash \alpha$' is read as 'α is true at w'.

Beall gives the following conditions as specifying the classical models:

S0 $w = w^*$

S1 $w \Vdash \alpha$ or $w^* \Vdash \neg\alpha$, for all $w \in N$

S2 $w \Vdash \neg\alpha$ iff $\neg(w^* \Vdash \alpha)$

S3 $w \Vdash \alpha \vee \beta$ iff $w \Vdash \alpha$ or $w \Vdash \beta$

S4 $w \Vdash \alpha \wedge \beta$ iff $w \Vdash \alpha$ and $w \Vdash \beta$

It is S0 which ensures classicality and, in this context, combines with S2 to make negation extensional. If S0 is dropped, then negation will cease to be extensional. But sticking with the current framework, and S0, we say that a sentence, α, is 'verified in a model' just if, in that model, $@ \Vdash \alpha$ and, for some set of sentences, Σ, $w \Vdash \Sigma$ just if $w \Vdash \beta$, for all $\beta \in \Sigma$. Let Σ be a set of sentences and α any sentence, we can then define, what Beall calls '*CPL** validity' as follows:

*CPL** validity: $\Sigma \vDash \alpha$ just if, for all classical models, if $@ \Vdash \Sigma$, then $@ \Vdash \alpha$

So, on this definition, validity is defined only in terms of the actual world, @, of each model. This picture is classical, and so Beall does not endorse it. In particular, he must give up the classicality constraint S0, since it does not allow for dialetheia. For example, suppose *a* is a dialetheia at a world, *w*, so we have $w \Vdash \alpha$ and $w \Vdash \neg\alpha$, then by the clause for negation, S2, we have $w \Vdash \alpha$ and $\neg(w^* \Vdash \alpha)$ (Beall call's w^* the 'star mate' of *w*). But then, by the classical constrain S0, we have $w \Vdash \alpha$ and $\neg(w \Vdash \alpha)$, which is impossible. One might wonder why this is impossible, since, for example, Priest has no problem with accepting contradictions of this kind. The reason is that Beall's metatheory is classical, so although we can have, on Beall's view $w \Vdash \alpha$ and $w \Vdash \neg\alpha$ (for some *w* and α), we cannot have $w \Vdash \alpha$ and $\neg(w \Vdash \alpha)$, since this would introduce inconsistency (and triviality) into the metatheory. In other words, sentences can be both true at a world and false at that world, but no sentence can be both true at a world and untrue at that world. This is a problematic feature of the semantics, it seems to me, and makes the picture undesirably artificial (I have also discussed this issue in the context of Priest view in Chapter 4, Section 5.2.1). One way to put the issue is as follows: truth, for Beall, commutes with negation. That is, the negation of a sentence, α, is true, if, and only if, α is

untrue. This is an important feature of the relationship between truth and negation. On the other hand, truth-in-a-model and negation do not commute. That is, the following does not hold:

$w \Vdash \neg\alpha$ iff $\neg(w \Vdash \alpha)$

So truth-in-a-model and '¬' interact very differently to truth and negation, and so the plausibility of the former as *truth*-in-a-model, and the latter as negation come into question. I suspect Beall's response to this would be to claim that this is only a problematic artificiality on the assumption that the semantics we're giving for our model language are intended to model those of our real language. Beall rejects this assumption, and takes the semantics to be 'merely heuristic', intending to give a theory modelling the real semantics of natural language later.

My counter to this is similar to the one made in Chapter 2 of this thesis in the context of defining revenge, and in Section 1.4 of the present chapter, discussing revenge on Beall's theory in particular. We need to be given this further theory if this response from Beall is to be adequate, so that we can see that the same problem will not arise in this new context.

The problem seems to be fairly general for dialetheist theories which take, as their metatheory, classical *ZF*. There can be sentences of the metatheory which say of an object theory sentence, *a*, that it is both true and false, but if we have a principle which takes us from the falsity of *a* to the negation of the metatheoretic sentence asserting the truth of *a*, then we have a contradiction in the metatheory. So we cannot have such principles, on pain of triviality, and this is apt to make the semantics artificial, assuming the view is committed to the transparency of the truth predicate (and so the equivalency of falsity with untruth) in the object language. It may be that Beall's yet-to-be-provided model of the semantics of natural language avoids this issue, but he must actually provide the model if this is to be convincing.

So, giving up S0 gives us the non-classical picture Beall wants. We keep each of S1-S4, taking them to give us the *LP** models. We then redefine (only slightly) validity as follows, again where $w \Vdash \Sigma$ iff $w \Vdash \beta$ for all $\beta \in \Sigma$

*LP** validity: $\Sigma \vDash \alpha$ just if, for all *LP** models, if $@ \Vdash \Sigma$, then $@ \Vdash \alpha$

This new, non-classical, logic allows for there to be dialetheia at a world, *w*, in cases where *w*'s star mate, *w**, is such that ¬(*w** ⊩ *α*) ∧ ¬(*w** ⊩ ¬*α*). In other words, *α* can be a dialetheia at a normal world, *w*, just if *w** is a world at which LEM fails. Given S1, this means *w** must be abnormal.

The resulting logic is *LP*, familiar from Chapters 3 and 4. One fact which Beall takes to be important is that, if *α* follows from Σ by the definition of validity *LP**, then *α* follows from Σ by the definition *CPL**. This follows because any classical model is an *LP** model (since, if a model meets S0-S4, *a fortiori*, it meets S1-S4). It is also true, though Beall refers the reader to Priest's (1979) for a proof, that *LP* and classical logic share the same logical truths. That classical logic is an extension, in the sense, of the 'real' logic *LP* is something which Beall takes to hold philosophical significance.

Next Beall demonstrates how predicates and quantifiers are introduced. The exact details need not concern us here, so I refer the reader to Beall's (2009, pp.11-12) for these. The truth predicate, *T*<*x*>, is introduced via the transparency principle:

Transparency: Let *β* be any sentence in which *α* occurs. Then the result of substituting *T*<*α*> for any occurrence of *α* in *β* has the same semantic status as *β*.

This is equivalent, on Beall's view, to what he calls 'Weak Transparency':

Weak Transparency: *w* ⊩ *T*<*α*> iff *w* ⊩ *α*, for all *α* and *w* ∈ *W*

This is taken as a constraint on the admissible models of *LP**, and so all *LP** models, by definition, satisfy Weak Transparency and, therefore, Transparency. In the current set-up, where *LP* contains only the material conditional, various desirable principles, which one might take to be constitutive of the notion of conditional, such as *modus ponens*, fail. So, as things stand, the logic is inadequate: a suitable conditional must be added. When this is done, Beall says, Weak Transparency no longer entails Transparency. The present, conditional-free, theory, which is essentially quantified *LP* with the addition of a transparent truth predicate, Beall calls *LPTT*. 18-24

Beall suggests two questions might arise about this set-up. The first asks whether there are any non-trivial models of the system. The second asks whether there are any 'natural' models of it, where 'natural' is understood to mean 'has a consistent, indeed classical, base language', where the base language is the truth-free fragment of the language. The answer to both questions is 'yes', and he lays-out, and discusses, these result in an appendix, (2009, pp.18-24). I follow Beall again in returning to the philosophical issues before finally giving Beall's account of the conditional.

5.3 Merely Semantic Dialetheism

To return to Beall's metaphorical story with which he motivates his view, we could, if we had the right capacities, describe the world entirely without truth. That is, the truth-free base language, 'in principle', suffices for a complete description of the world. The reason we invoke a transparent truth predicate is not that is picks out some feature of the world which is missed by the base language, but rather that it allows us to overcome, in a sense, our finitude. It doesn't generate 'new claims' about the world, but, rather, 'reveals' claims about the world that we couldn't otherwise express due to our various limitations. According to Beall, a being without such limitation could do without a truth predicate (2009, p.14).

Even so, because of the grammatical features of languages like English, new claims arise anyway which are truth-ineliminable, such as the familiar liar sentences as well as truth-tellers like *TT*:

(*TT*) The sentence *TT* is true.

Because of the way this sentence is constructed, the occurrence of the truth predicate cannot be eliminated by the substitution rules for the truth predicate. *TT* differs from the liar, however, in not providing a direct argument to inconsistency, since it can quite consistently be either true or false; it's just that nothing would seem to decide which. Beall opts for what he takes to be the simplest approach to such sentences, which is to treat all such truth-ineliminable sentences as dialetheia, though he open to the possibility of treating such sentences asymmetrically (as either just true or just false). The

position is actually slightly awkward. Curry sentences, too, are truth-ineliminable, but Beall cannot, if he is to avoid triviality, treat this as a dialetheia. Acknowledging this point, Beall restricts his claim, that all truth-ineliminable sentences are dialetheia, to the conditional-free fragment of the language (2009, p.34). He does not give an account of how we are to decide which truth-ineliminable sentences are gluts in the language generally. But, even in the conditional-free part of the language, we can still construct sentences like "This sentence is true and JC Beall is a fried egg". If this is treated as a dialetheia, then a base-language contradiction follows, that JC Beall both is and is not a fried egg. If all such sentences are treated as dialetheia, then triviality follows.[8] So, in fact, it is unclear what is available to Beall in terms of a general account of which truth-ineliminable sentences are dialetheia, such that it vindicates the claim that the truth-teller is a dialetheia, without committing him to triviality via Curry (or Curry-like) paradoxes.

So, on Beall's view, the truth-free fragment of our language, the base language, is entirely consistent and classical. When we add our generalising expressive device, $T\langle x\rangle$, it brings with it dialetheic spandrels: truth-ineliminable sentences which are both true and false. But this, according to Beall, is the *only* way dialetheia arise. In this sense, dialetheia are a 'merely semantic' phenomenon.

The way to make this more precise, according to Beall, is to let $\mathcal{L}$ be our base language and model the idea that dialetheia are 'merely semantic' by stipulatively restricting our class of *LP** models to only those such that $\neg(@ \Vdash \alpha \wedge \neg\alpha)$, for any $\alpha \in \mathcal{L}$. The effect of this is that there are no models containing base-language dialetheia. One advantage of this feature, which Beall highlights (2009, pp.16-17), is that, though they are not, of course, valid, both disjunctive syllogism and explosion are without counterexample in the base language. We might call this, using Priest's terminology (2006, p.110) 'quasi-valid'. The advantage is that one can safely use these inference principles in when reasoning in the base language.

So this is Beall's 'merely semantic dialetheism': our truth-free base language is perfectly consistent and classical, and were it not for our limitations, this

[8] This point was made by Ben Burgis to me in conversation, and in his unpublished manuscript, *Does Beall's Version of Dialetheism Avoid Triviality?*.

language would have been sufficient. But because of our limitations, we had to introduce a transparent truth predicate, which brings with it inevitable dialetheic spandrels. But because this is how these spandrels are introduced, dialetheia are 'merely semantic' and contradiction does not 'infect' the rest of our (consistent) language.

It seems to me a serious problem arises here for Beall's view. One of the many models of *LP** is the trivial model, in which every sentence is both true and untrue. Beall rejects this model of the logic. The reason he gives is its lack of usefulness, since the "trivial model cannot invalidate anything that isn't already invalidated." (2009, p.64 n.10). This is a peculiar reason to reject the legitimacy of a model. Even if we admit that the trivial model does no work on its own, it doesn't follow that we ought to deny its very existence. Some appeal to parsimony might be made here, that we should accept the existence of as few models as we can get away with (in some sense of 'get away with'), but this would be difficult to substantiate, given the loss of theoretical parsimony from adding restrictions ruling out what would otherwise be perfectly legitimate models of the logic. It also doesn't follow from the fact that a model doesn't invalidate anything that it does no theoretical work. Models of the logic give us a guide to what is logically possible, on that logic. The existence of the trivial model, then, tells us that the trivial situation is logically possible in *LP**. So it is not theoretically inert: far from it, as we shall see in a moment.

Perhaps the worst problem facing this reason for rejecting the existence of the trivial model is that it massively over-generalises. No counterexample to any inference principle, at least none of which I am aware, and none that Beall mentions, depends for its efficacy on the particular, identifying features of a single model. Simplifying to the single-premise case (and to the cases where counter-examples involve a single world – things are more complicated in some cases, as can be seen in the discussion of Beall's conditional, below), a counter example to an inference principle $A \vDash B$ is a world, w, at which A holds but B fails to hold. For any such world, we can find a different one in which A and B take the same values, but which differs with regard to the truth value of some further sentence, C, which is not relevant to the inference principle in question. If this is right, then *every* model is such that it cannot invalidate anything which

is not already invalidated, since excising that model from the theory does not, on its own, affect which inference principles are invalid.

On the other hand, it is extremely important for Beall that there be no trivial model. Recall from section 1.1.1 the modal law of non-contradiction, to which Beall is committed:

(*MLNC*) $\neg\Diamond(P \wedge \neg P)$

For any merely possible contradiction, α, we have $\Diamond(\alpha \wedge \neg\alpha)$ and, by *MLNC*, $\neg\Diamond(\alpha \wedge \neg\alpha)$. So any merely possible contradiction delivers the actual contradiction that it is both possible and impossible. Importantly, if the α is a sentence of the base language, then $\Diamond(\alpha \wedge \neg\alpha) \wedge \neg\Diamond(\alpha \wedge \neg\alpha)$, too, is base language, because whether a sentence is in the base language is a matter only of whether the sentence contains the transparent truth predicate $T\langle x\rangle$. So, if Beall's theory allows for a trivial model, then since every sentence, including every sentence of the base language, is both true and untrue there, then for every sentence β, of the base language, we have $\Diamond(\beta \wedge \neg\beta)$ and (by *MLNC*) $\neg\Diamond(\beta \wedge \neg\beta)$. The existence of a trivial model, then, would (assuming its accessibility from the actual world) refute Beall's merely semantic dialetheism, by swamping the base language with inconsistency.

The point, moreover, extends well beyond the trivial model. If there is *any model at all* containing a base language contradiction, we can run the same argument to generate an actual base-language contradiction. The principle *MLNC* effectively says that no contradiction is logically possible. But since many contradictions are logically possible (indeed, actual), *MLNC* is a dialetheia: all contradictions are logically impossible, but some are possible too. However, the argument given above means that Beall is committed to every base language contradiction being logically impossible *only*. Since Beall is a dialetheist, it is difficult to see what could motivate this claim.

As an aside, one might think that Beall cannot even express that certain things are logically impossible *only*, because of the just false problem, since for something to be impossible *only* is, presumably, for it to be just false in every

model (or world). So, unless Beall had some consistent notion of ‘just false’, ‘just impossible’ could not be expressed without including some possible things too. This is not so. Since Beall’s metalanguage is classical, he can, in the metalanguage, restrict the admissible models of the logic to those free of base-language contradiction. This is exactly the effect of the restriction above that, with $\mathcal{L}$ the truth-free fragment of the language, we restrict the admissible models to those such that $\neg(@ \Vdash \alpha \wedge \neg\alpha)$, for any $\alpha \in \mathcal{L}$. The consistency of this restriction on the base language, then, is guaranteed by the consistency of the metatheory.

So Beall is able, though only via the metatheory, to restrict the admissible models of the logic and so force consistency on the base language. The problem, then, is not with his ability to restrict the theory in this way, but in motivating the restriction. This, it seems to me, is the important upshot of the debate between Beall (2011) and Eldridge-Smith (2011, 2012) concerning the Pinocchio Paradox (though it is not made clear in the exchange itself).

Eldridge-Smith (2011) tells a story about a puppet, Pinocchio, whose nose grows if and only if he speaks a falsehood. Pinocchio utters the sentence ‘My nose is growing’ and, according to the story, the world explodes in triviality. The problem this is supposed to pose for Beall arises from the fact that the contradictory sentence ‘My nose is growing’ is entirely base-language. It is true that the principle specifying the behaviour of Pinocchio’s nose contains the falsity predicate, but this doesn’t matter, since it doesn’t affect the status of ‘My nose is growing’ as a sentence of the base language.

Beall’s response (2011) invokes the more familiar Barber paradox for comparison. In the story of the Barber paradox (first described by Russell (1986) to describe the Russell Paradox), there is a barber who shaves all and only those people in his town who do not shave themselves, and we consider whether this barber shaves himself (he does if and only if he doesn’t). Beall asks why dialetheists are not committed to contradictions involving barbers, and answers that it’s because the story is a fiction: the correct conclusion to draw from the barber paradox is that there are no such barbers. This is, of course, correct, and no one without an antecedent commitment to a comprehension principle for

barbers is likely to disagree. He thinks the Pinocchio Paradox generates no further problem.

For this response to be adequate, we would have to assume that, in order to cause problems for Beall's view, Pinocchio (or the barber, for that matter) would have to actually exist. But, as we can see from the arguments above employing *MLNC*, this is not so. All that's required for Pinocchio, or the barber, to cause inconsistency in Beall's base language is that they be merely logically possible. We can agree with Beall that Pinocchio does not actually exist and, further, we could accept that he is physically impossible; perhaps (depending on the account) we might accept that he is metaphysically impossible, biologically impossible, or whatever. But so long as there is some sense of possibility on which Pinocchio is possible and for which we have a corresponding version of *MLNC*, the base language must be inconsistent.

The same is so for the Barber. There are, presumably, no actual barbers of the kind described in the story. As with Pinocchio, such barbers may be physically, even metaphysically, impossible. But so long as there is a sense of possibility, no matter how weak, on which the barber is possible and *MLNC* holds, base language contradiction follows from that possible barber too.

As an aside, we should be slightly careful about what exactly we mean when we say that Pinocchio is logically impossible. What is impossible is for there to be a puppet (or any other entity) which has a nose obeying the principle that it grows if and only if it speaks a falsehood, and which utters the sentence 'My nose is growing'. So talking puppets called 'Pinocchio', including ones which utter 'My nose is growing' are, presumably, logically possible, so long as their noses don't obey the principle above. Similarly, puppets whose respective noses obey the principle that they grow if and only if they speak falsehoods are also, presumably, logically possible, so long as they keep quiet. Or, at least, so long as they don't utter the sentence 'My nose is growing'. What's impossible is the combination of the principle connecting nose growth to the speaking of falsehoods and the 'liar sentence' uttered.

One difference between the two cases is that the Barber Paradox is structurally isomorphic to Russell's paradox, whereas the Pinocchio Paradox is structurally

isomorphic to the Liar Paradox. Since Beall endorses a classical picture of set theory (2009, p.112), he endorses a non-existence solution to the Russell Paradox. But, of course, he endorses a dialetheic solution to the Liar Paradox. So the Pinocchio Paradox has the advantage that is shares a structure with a paradox for which Beall accepts the dialetheic conclusion.

The Pinocchio Paradox involves a puppet with various features which we don't think are realised in the actual world and which, as I have said, we can accept are impossible to realise in a number of senses (physically, perhaps metaphysically, perhaps others), but these do not speak to Pinocchio's *logical* possibility. If we're to accept as plausible the claim that Pinocchio is logically impossible (only), what we need is some logical asymmetry between the liar paradox and the Pinocchio paradox which would explain why one is necessarily dialetheic, but the other impossible only. But since they are structurally isomorphic, it is extremely unclear what this could be. Those endorsing a classical view of logic will think that Pinocchio is impossible, as they will the liar, on the basis of the more general claim that there are no contradictions. Since Beall does not endorse this more general claim, he can't reject the possibility of Pinocchio on this basis.

So the Pinocchio Paradox is just a concrete instance of a more general problem, which is that Beall needs to provide some motivation for taking the base language to be completely consistent, given the existence of principles like *MLNC* which deliver base language inconsistency from merely logically possible base language contradictions. The value of the Pinocchio paradox is simply in providing a particular case where it seems strikingly implausible that Beall, a dialetheist, should think it ruled-out as impossible only, on the grounds of logic.

One asymmetry to which Beall might wish to appeal is that the putative contradiction in the Pinocchio case is base language, and so generates base language inconsistency, which Beall rejects. But in the present context, this is entirely question-begging. The very thing at issue is whether Beall has given us good reason for supposing the base language to be consistent.

One might, instead, look for an asymmetry, or independent motivation for the consistency of the base language, in Beall's picturesque story which motivates

his deflationism and his dialetheism, given in this chapter. But this doesn't provide what we want either. Beall's story is that we should see dialetheia as arising from our transparent truth device, which we introduce to escape our finitude and express certain generalisations which we otherwise couldn't, and which comes with inevitable spandrels like the liar. But, given our capacity to introduce inconsistent predicates which generate dialetheia, we are stuck with inconsistency and we are thereby stuck (on the assumption that we wish to reject triviality) with paraconsistent logic. Along with this paraconsistent logic, which allows for inconsistency, comes a broader understanding of logical possibility than we get on the usual classical picture. There is nothing in this logic itself to preclude inconsistent entities like Pinocchio and so, via *MLNC*, the base language contains certain contradictions. We nonetheless think of our dialetheism as arising from our understanding of transparent truth, in exactly the way Beall describes, it's just that the appropriate logic allows for the logical possibility of base language contradictions (there's nothing *logically* wrong with them), and so the base language ends up, perhaps surprisingly, to allow in dialetheia through the back door. So Beall's commitment to base-language-consistency is not justified by the more general motivations he gives for his view. We can accept the basic story and still have plausible account of how base language inconsistency arises.

The general problem, again, is that a part of Beall's view on which he places special emphasis is the restriction of dialetheia to the semantic fragment of the language and the consistency (and classicality) of the base language. That Beall's logic has a trivial model threatens this claim, in the ways described, so Beall shuns the existence of this model. His reasons for doing so, I have argued, are not good. But the problem is worse, since he must rule out the existence of every model containing base language contradiction: all such models (or worlds) are logically impossible only. He has not given a motivation for this, and it is difficult to see how he could. Pinocchio provides one intuitive case which seems like it clearly shouldn't, in virtue of its similarity to the liar, be impossible only. Moreover, there seems a perfectly plausible story to tell, in complete accord with Beall's basic motivational story for his view, of how the base language ends up inconsistent.

Beall might, perhaps, accept this and decide instead that the consistent fragment of our language is the truth-free *non-modal* part of our language; that is, the fragment not containing either a truth predicate or any modal notions. This is certainly one option and I know of no reason that inconsistency should spread here. It makes the view less interesting, however, in part because Beall's conditional is, as we shall see, modal, and so this fragment of the language would be conditional-free.

5.4 Beall's Conditional

As I have said, the theory previously sketched, *LPTT* (*LP* plus a transparent truth predicate), fails to support important inference principles concerning the conditional such as *modus ponens*. So the conditional (if this is not a misnomer) of *LPTT* is inadequate and must be extended with a suitable conditional supporting principles like *modus ponens*.

As in the case of Priest's view (see my discussion in Section 3.2.0 of Chapter 3) one of the crucial constraints on a conditional is that it avoid the threat of triviality from Curry's paradox. Which, in the present context, is to say, it must avoid any form of contraction:

$(\alpha \wedge (\alpha \rightarrow \beta)) \rightarrow \beta$
$(\alpha \rightarrow (\alpha \rightarrow \beta)) \rightarrow (\alpha \rightarrow \beta)$
$\alpha \rightarrow (\alpha \rightarrow \beta) \vDash \alpha \rightarrow \beta$

If any of these hold for '$\rightarrow$', triviality follows via the curry paradox. It is therefore a constraint on a conditional being suitable that each of these fail. A further constraint is that the conditional satisfy identity. That is, we must have $\vDash \alpha \rightarrow \alpha$. The reason, aside from its being independently compelling, is that it is from indentity, combined with the substitution rules by which we introduce our truth predicate, that we obtain the crucial principles for truth, *capture* and *release*:

Capture: $\alpha \rightarrow T\langle\alpha\rangle$
Release: $T\langle\alpha\rangle \rightarrow \alpha$

So Beall stipulates that a conditional is suitable just if it detaches, validates identity and avoids any triviality-inducing principles of contraction. Beall offers a bifurcated account of the conditional which gives it different truth conditionals depending on the sort of world in which it occurs. One principle is given for conditionals at normal worlds, and a separate one for abnormal worlds. We expand the structures already given with a ternary relation on worlds R , which we think of in terms of accessibility, such that $R_{w,w',w''}$ holds just if the pair $\langle w',w''\rangle$ is accessible under the relation from w (Beall calls the pair 'w-accessible'). The reader can consult Chapter 2 of Beall's (2009) for a fuller treatment than I give here.

Normal. Where $w \in N$ and $w' \in W$,

$w \Vdash \alpha \rightarrow \beta$ iff for or any w' $\in W$, If $w' \Vdash \alpha$, then $w' \Vdash \beta$

Abnormal. Where $w \in W - N$ and $w',w'' \in W$

$w \Vdash \alpha \rightarrow \beta$ iff, for any w-accessible $\langle w',w''\rangle$, if $w' \Vdash \alpha$, then $w'' \Vdash \beta$

So, if a conditional, $\alpha \rightarrow \beta$, occurs at a normal world, it is true just if there is *no world* (normal or abnormal) at which α holds, but β doesn't. But, if the conditional occurs at an abnormal world, then we are constrained by what is accessible from that world, so the conditional $\alpha \rightarrow \beta$ will hold at an abnormal world just is there is *no pair of worlds accessible from it* where α holds at one, but β doesn't at the other.

The conditional has the advantage that *modus ponens* holds of it, as well as identity, and that contradiction fails. Beall calls the resulting logic *BXTT* (*BX* plus transparent truth), and gives an axiomatic system for it which can be shown to be complete (Beall presents the axiom system in his (2009, pp.31-32) and refers readers to (Priest and Sylvan, 1992) for the completeness proof).

A final formal issue is how Beall treats validity. This is, ordinarily, as is intuitive, defined in terms of the preservation of truth. This is not how Beall defines validity. The reason is that, if we introduce a validity predicate, *Val*, into our language and define it in terms of necessary preservation of truth, this gives us:

V1. $Val(\langle\alpha\rangle,\langle\beta\rangle) \rightarrow (T\langle\alpha\rangle \rightarrow T\langle\beta\rangle)$

Where '→' is a detachable conditional, perhaps the one defined above, this gives us, along with principles about conjunction valid in *BXTT*:

$Val(<\alpha \wedge (\alpha \rightarrow \beta)>,<\beta>)$

Which, gives us:

$(\alpha \wedge (\alpha \rightarrow \beta)) \rightarrow \beta$

This, being a form of contraction, gives us triviality by Curry's paradox. So Beall must give up the connection between validity and truth preservation, at least in any form strong enough to deliver V1.

Beall settles on the following principle (which he restricts to the single-premise case for simplicity) as a characterisation of the sense in which, on his view, validity relates to the preservation of truth:

V2. $Val(<\alpha>,<\beta>) \rightarrow (\alpha \supset \beta)$

Where '⊃' is a material 'conditional'. This still delivers a form of contraction, namely:

$(\alpha \wedge (\alpha \rightarrow \beta)) \supset \beta$

But, since ⊃ does not detach, we can't derive triviality from Curry sentences, this form of contraction is benign. For the same reason that this principle is harmless (it does not detach) V2 is an extremely weak connection between validity and truth-preservation. In particular, it means that one cannot proceed from a valid argument and the truth of the argument's premises to the truth of the argument's conclusion.

Beall's response to this apparent problem is to downplay the importance of being able to detach the conclusion of a sound argument. The 'utility' of valid reasoning is to ensure that our reasoning has the right sort of 'steps' (2009, p.36), specifically that each step must go from truth to truth, from untruth to truth, or from untruth to untruth. This, that each step is of the proper kind, Beall think, is delivered by V2. That this is the point of reasoning is contentious and, on the face of it, wrong. Surely, one might think, the point of valid reasoning is that I can begin with true premises and deduce, by that reasoning,

the truth of the conclusion. If it is true that I am in Glasgow and that, if I am in Glasgow, I am in Scotland, the point of reasoning validly from these premises is, on the face of it, that I can obtain the truth of the conclusion that I am in Scotland. Beall's claim that the point of valid reasoning is merely a matter of having the right 'steps' needs more argument than is given. One might think by emphasising the importance of the right steps, Beall is advocating a proof-theoretic account of logical consequence. But he does not do this, so it is unclear what the 'steps' delivered by V2 really have to do with the utility of valid reasoning, especially given the fact that he admits the 'proper steps' may take one from true premises to an untrue conclusion (2009, p.36).

In fact, things are worse for V2, because Beall's conditional, '→', contraposes (2009, p.32). So, from the contraposability of the conditional and the definition of '⊃', V2 immediately gives us V2*:

V2*. $\neg(\neg\alpha \vee \beta) \rightarrow \neg Val(\langle\alpha\rangle,\langle\beta\rangle)$

By the De Morgan equivalencies (and double negation elimination), this is equivalent to V2**:

V2**. $(\alpha \wedge \neg\beta) \rightarrow \neg Val(\langle\alpha\rangle,\langle\beta\rangle)$

So, still restricted to the single-premise case, if we can show that $\alpha \wedge \neg\beta$, then since V2 delivers V2** by principles which hold in *BXTT*, the inference from α to $\neg\beta$ is invalid. This makes Beall's account of validity inconsistent, since we can show this for lots of valid inferences. To take the simplest example, take the inference principle *reflexivity*: $R \vDash R$. Consider a simple liar sentence, L, which is such that $T\langle L\rangle \wedge \neg T\langle L\rangle$ and, on Beall's view, is also such that $L \wedge \neg L$. By V2**, we have $\neg Val(\langle L\rangle,\langle L\rangle)$, and so a counterexample to reflexivity. Since, presumably, reflexivity is valid too, we have $Val(\langle L\rangle,\langle L\rangle) \wedge \neg Val(\langle L\rangle,\langle L\rangle)$. So, on Beall's account, just as on Priest's, validity is inconsistent. The point extends to a great many other inferences, of course, beyond reflexivity. The reader can consult Chapter 4 of this thesis to see how this is done.

The problem for Beall here is more serious than for Priest, in at least a couple of ways. The first is that dialetheic validity is a consequence (and cost) of Priest's paraconsistent metatheory, which has other advantages, such as the lack of a

real object/metalanguage distinction. To accept the inconsistency of validity would be to end up with the costs of a dialetheic metatheory without the benefits. Another disadvantage is that, I take it, the reason that validity is generally understood as a semantic notion is that it is defined in terms of (the necessary preservation of) truth. Since Beall explicitly does not define validity this way (if it is defined at all, of which more below), there seems little reason to think the notion semantic. This being so, contradictions like $Val(\langle L\rangle,\langle L\rangle) \wedge \neg Val(\langle L\rangle,\langle L\rangle)$ are in real danger of being *non-semantic* contradictions. The existence of which would refute Beall's 'merely semantic' brand of dialetheism, independent of their intrinsic implausibility.

So Beall rejects that validity is definable in terms of the preservation of truth, accepting only the weak connection between these notions given by V2. But he can't accept even this weakened principle, without making his account of validity inconsistent and, unless he can make plausible that his notion of validity is still semantic, giving up his merely semantic form of dialetheism.

A further restriction on the addition of a validity predicate $Val(\langle x\rangle,\langle y\rangle)$ to the object language is the validity Curry paradox, which Beall himself has discussed in his more recent work *Two Flavors of Curry Paradox* (2011). This paradox arises from a sentence which, in effect, says of itself that the argument from it to triviality is valid:

C: $Val(\langle C\rangle,\langle \bot\rangle)$

The argument to triviality exactly mirrors that of the more usual conditional Curry paradox. There are a number of principles governing *Val* which one might give up to avoid the problem. One is structural contraction, the rejection of which would mirror Beall's rejection of the parallel contraction principle in the conditional Curry case. This would require a substructural logic, rather than the logic *BXTT*, which Beall endorses in his (2009). Other options are available, and the problem is not unique to Beall's view. I mention it only as a further potential source of triviality, arising from the addition of *Val* to the language, and so as a further constraint on what principles *Val* may be taken to satisfy.

Since he doesn't take validity to be necessary preservation of truth, he immediately faces the question of what he *does* think validity is. He does not

take a firm view of the matter, but offers three potential options, each of which, despite problems, he would be happy to endorse.

5.5 Accounts of Validity

The first option is not to offer a *definition* of validity at all, but to take the notion as primitive. Though, strictly speaking, there is no definition of the notion, on this option, we may still give principles which it obeys. Presumably he intends V2 to be one of these, as well as, for example:

$Val(<\alpha>,<\beta>) \rightarrow (\alpha \rightarrow \beta)$

Since the '→' here is Beall's detachable conditional, which delivers contraction if accepted as full general, this could only hold in certain fragments of the language. A further constraint one might wish to put on validity is normative:

One ought to reject the validity of a given argument if one accepts that it's possible for the given premises to be true but the conclusion false.(2009, p.38)

It is surprising that Beall suggests this principle, since he has just accepted that one can have valid arguments with true premises and an untrue (which is equivalent, for him, to false) conclusion. This would commit him to the view that one should reject some valid arguments as invalid. Since those arguments are valid, presumably, one also should not reject those arguments as invalid. So, it may be that this principle commits Beall to there being things such that some things ought to be rejected, though it's not the case that they ought to be rejected (or, perhaps, just that they ought to be accepted and rejected). Priest accepts the possibility of this sort of contradiction (2006, p.274), so there is some precedent for the view, but things are more difficult in Beall's case. If, and this may depend on details yet-to-be-filled-in, there is a genuine dialetheia here, it does not appear to be a semantic one, since acceptance and rejection are not, on the face of it, semantic notions. This sort of dialetheia would undermine Beall's merely semantic account of dialetheia. Even if we do not have an explicit dialetheia involving rejection, i.e. there are no principles which we both reject and do not, rejecting the validity of a principle must commit us

(assuming excluded middle) to its invalidity. If so, then some valid principles are also invalid, and we are back with the same problems as with V2.

A further problem which Beall highlights is epistemological. If validity is primitive, how do we know which arguments are valid and, relatedly, how do we know which relation in the language (if any), is the validity relation? Beall admits this is a serious problem, and says that he has no answer to it. The sting of the objection, as a particular problem for the view on which validity is primitive, however is lessened by the fact that, according to Beall, the problem is not peculiar to this account. Were we to define validity as necessary preservation of truth, we would still face questions such as "How do we know *that* is a 'real counterexample'?" (2009, p.38) But the cases are not exactly parallel here. In Beall's case, our notion, validity, is given no definition, and seemingly very few principles which govern its application. I have just suggested that the above principle, involving rejection, may well be unsuitable. We also, as I have argued, don't have V2. This being so, we have very little to go on in identifying which are the valid arguments, on this primitive account. If, on the other hand, we define validity as preservation of necessary truth, then whether, for example, something is to count as a 'real counterexample' is simply a matter of the truth values of the sentences contained in the counterexamples. How easy it is to decide this matter will, of course, depend on the sentences in question, but it's difficult to see how, without taking a position of fairly general epistemological scepticism, how this position could fail to be better-off, epistemologically speaking, than Beall's primitive account of validity.

The second option Beall calls a 'two-step account' of validity. On this strategy, one constructs a model language, using model theory, and defines validity for that language, in its metatheory, using the notion of truth-in-a-model. This is the first step. The second step is to 'extract' the valid argument forms so-defined and take them to be a definition of validity for our real language. This strategy has in common with the primitive one that validity is not defined in terms of truth (though it is defined in terms of truth-in-a-model. Beall points to two problems for this view. Firstly, we face similar epistemological problems as in the case of the primitive approach: "Why *that* formal language versus *that* on[?]"(2009, p.40) Secondly, we require a precise account of 'argument forms', which is, he admits, not an easy matter.

Another problem for this account is that, presumably, our account of validity no longer involves the introduction of a validity predicate characterising validity in and for our object language. On this view, then, the model language cannot characterise its own validity relation. Natural language, on the other hand, (presumably) can.

A final option is to offer a 'partial definition' of validity. On this account, we give only sufficient, rather than necessary, conditions for validity. For instance, we may give the following condition as sufficient for validity:

V3. $(\alpha \rightarrow \beta) \rightarrow Val(\langle\alpha\rangle,\langle\beta\rangle)$

Beall says "This might not be everything that one would want from an account of validity, but it may be all that we need." (2009, p.41) The onus is on Beall, it seems to me, to give arguments for why we should expect this to give us all we want from an account of validity. One might wonder what the material difference is between this approach and the primitive one. In neither case do we give a full definition of validity, and in both cases it is in our interests to endorse as many and as strong principles as we can get away with, to give us the best grip possible on validity, despite the lack of a definition. Whether we count these principles as a partial definition of validity, or merely as constraints on the use of a primitive notion, seems like a distinction without a difference.

5.6 Some General Considerations about Metatheoretically Consistent Dialetheism

Though this thesis is about the revenge problem in particular, it will be helpful in comparing the relative strengths of different varieties of dialetheism to discuss some other of the issues these views face, before going on, in the final chapter to discuss revenge problems for the view. This chapter has characterised Beall's particular version of metatheoretically consistent dialetheism, and has given some problems for the view thus characterised. But Beall's view is not the only available, so this section discusses a couple of more general differences between metatheoretically consistent, and metatheoretically inconsistent, dialetheism. Since, again, my focus in this thesis is revenge, I do not mean the

discussion in this section to be definitive, by any means, but rather to help bring out the differences between the different kinds of dialetheism.

I begin, though, with a final issue concerning Beall's merely semantic dialetheism. Priest's more thoroughgoing dialetheism allows for non-semantic dialetheia. Indeed, Priest defends dialetheic accounts of set theory, vagueness, change and motion, indefinite extensibility, inconsistent legal systems and others. One effect this has is to make his version of dialetheism, arguably, more radical. Another effect is that, on the assumption that these applications are successful, they provide further support for the view. If one accepts the existence of true contradictions, and, this done, discovers that accepting such things provides solutions, without further significant costs, to long-standing, intractable philosophical problems, it would seem reasonable to increase ones confidence in the existence of dialetheia.

Since each of the applications of dialetheism mentioned (plausibly) involves accepting the existence of non-semantic dialetheia, none of them are available to someone, like Beall, who endorses merely semantic dialetheism. To take an example, Priest endorses a dialetheist account of vagueness, where the familiar Sorites paradoxes are taken to involve dialetheia at points of 'overlap' between the correct application of a predicate and its negation. The details of the account need not concern us here. If Priest can make plausible a dialetheist account of vagueness, without significant extra cost beyond the inconsistency already accepted, then his version of dialetheism has gained significant additional support by solving an extremely intractable philosophical problem. Paracomplete solutions to the paradoxes, like Field's, can also gain this sort of further support by treating the vague, borderline cases as 'gaps'. This sort of support is denied to a view like Beall's.

Perhaps the most significant objection to such non-classical solutions to puzzles of vagueness is that they are non-classical: they require one to give up classical logic. But this objection, obviously, has much less force against someone who endorses a non-classical logic for independent, in this case liar-paradoxical, reasons. On the other hand, defending a classical account of vagueness comes with its own costs. For instance, in Williamson's *epistemicist* account of vagueness, every 'vague' predicate has a precise, but in-principle-unknowable

cut-off point; for example, there is a unique, decisive hair which, once lost, changes one from precisely non-bald to precisely bald. On Beall's own, rough sketch of a classical account (2009, pp.107-110), complicated machinery is introduced appealing to 'positive predicates', 'atomic predicates' and 'overline mates'. Vague predicates involve 'gaps' between positive atomic predicates and their overline mates, though sharp, classical cut-offs still exist between the correct application of any predicate and its negation. The proper philosophical interpretation of these notions is extremely opaque, and it does not avoid the implausible consequence that vague predicates admit of sharp, classical cut-offs. This is not intended to be a definitive objection to these accounts, but a worry about them being endorsed by someone like Beall is that, by accepting non-classical logic on the basis of liar reasoning, one is already paying the central cost for non-classical solutions to the puzzles of vagueness, but additionally taking on the inevitable costs of preserving classical logic in those contexts. If it turns out that the costs to a non-classical logician of endorsing a classical account of puzzles like, for example, vagueness, are liable to outweigh the costs to that logician of simply extending the scope of their non-classicality, then views like Beall's merely semantic dialetheism are likely to be unstable.

Though it's a commitment of Beall's version of metatheoretically classical dialetheism that he not accept potential contradictions arising from vagueness, it is not a commitment of the view in general. Set theoretic dialetheia, on the other hand, are different. While it is not, strictly speaking, off-limits for a dialetheist to accept that there are set-theoretic dialetheia, but keep their metatheory classical, there is a tension in the view. If there are true contradictions in set theory, then (on the assumption that the theory isn't trivial) the right theory of the sets is paraconsistent. But one of the crucial roles for set theory, in the present context, is as our metatheory which allows us to characterise the features of our model, object language. The position, then, would be one in which the correct set theory is paraconsistent, but the best theory of sets for the (central) purpose of providing our metatheory, is classical, which, if not outright inconsistent, is bordering on incoherence. Supposing this position is not adopted, then this version of dialetheism cannot endorse the powerful argument's given by Priest (and discussed in Chapter 3 of this thesis) for (inconsistent) naïve set theory. This is a fairly significant loss, and a prima

facie disadvantage to this form of dialetheism, as compared with Priest's more thoroughgoing kind.

Another, related point, is that Priest's dialetheism allows for the collapse of the distinction between object theory and metatheory. This gives Priest's view the attractive feature that his object language is able to describe its own semantics, which, presumably, is something it shares with English and other natural languages. This is another bonus denied to dialetheists whose metatheory is classical.

A further issue concerns the universality of logic. If one thinks logic is universal, which is to say, there is a single correct logic for reasoning about any subject matter, then an asymmetry in logic between one's object theory and metatheory is problematic. The logic one takes to be correct, presumably, is the one which holds in the object theory. But if this logic is the correct one for reasoning about any subject matter, then it ought to be the correct one for reasoning about itself. In other words, if the logic of the object theory is the one true logic, it ought to be the logic of its own metatheory.

So, if one has a paraconsistent logic in the object theory, but classical logic in the metatheory, there is some pressure to be a logical *pluralist*. This is the view that there is a general notion of logical consequence, which admits of many, equally good, precisifications. So, relevant logic, classical logic and intuitionistic logic are each, in some sense, equally good. This is the view defended in, for example, *Logical Pluralism* (2006), by Beall and Restall. But the view is controversial; Stephen Read (2006) has argued that it is incoherent, for example. If one can make a good case for this view, and in particular, the claim that classical logic is the right one for the metatheory, this may not be problematic, and I don't take a view on whether it is or not here. But depending on one's view of these matters, the pressure on defenders of this view to adopt a pluralistic view of logic may be unwelcome.

One issue, in the other direction, is that dialetheists who have a paraconsistent metatheory are hostage to the fortunes of paraconsistent set theory, which is still in the early stages of development. It may yet turn out that paraconsistent set theory is too weak to give a proper account of the sets. If this is so, it will be

difficult to maintain it as the correct metatheory for a paraconsistent logic. On the other hand, whether this constitutes a reason, now, to prefer a classical metatheory is unclear. As I discussed in Chapter 3, some recent result by Zach Weber give us reason for tentative hope about the prospects of paraconsistent set theory.

I have argued, in Chapter 4 of this thesis, that dialetheists whose metatheory is paraconsistent must treat validity inconsistently, and that, from this, a revenge problem arises. One might think that this can be avoided by endorsing a classical metatheory. The situation is not completely straightforward. If one wants to introduce a validity predicate to characterise validity in, and for, the object language, then the possibility of that predicate behaving inconsistently is a live one. As I have argued, Beall's view, at least in the form presented, has this problem. Suitable restrictions could, of course, be placed on the predicate to stop this, but whether the resulting predicate would look anything like validity is a matter which can be reasonably disputed. One could also simply define validity in the metalanguage without including a validity predicate in the object language, and thus ensure the consistency of validity by the consistency of the metatheory. Whether this constitutes a serious advantage of this approach over metatheoretically paraconsistent dialetheism is unclear. The real problem arising from the inconsistency of validity, I argue (in Chapter 4), is that is gives rise to a revenge problem for the view. If, as I also argue in the next section, metatheoretically consistent dialetheism also has its revenge problems, then it's not obvious they are at an advantage on this score.

5.7 Chapter Conclusion

The purpose of this chapter has been to describe metatheoretically consistent dialetheism, especially as defended by JC Beall in his (2009). Beall's 'merely semantic' dialetheism was introduced and critically discussed, and more general remarks made about the differences between metatheoretically consistent dialetheism, on the one hand, and metatheoretically paraconsistent dialetheism, on the other.

Chapter 6: Getting Revenge on Metatheoretically Consistent Dialetheism

6.0 Introduction

I argued in the first chapter of this thesis that revenge comes in essentially to varieties. The first is *formal revenge*, which, roughly, involves the construction of a notion in a theory's metalanguage which, were it expressible in the theory's object language, would lead generate a liar paradox rendering some notion expressible in natural language inexpressible in the theory's object language. The second is *informal revenge*, for which, again roughly, we bypass the theory's metalanguage and find a notion expressible in natural language which, were it expressible in the theory's object language, would generate a liar paradox, rendering some notion expressible in natural language inexpressible in the object language. Metatheoretically consistent dialetheism suffers from both sorts of revenge problem. I split this section into two subsections, the first concerning formal revenge, the second concerning informal revenge.

6.1 Formal Revenge for Metatheoretically Consistent Dialetheism

The recipe for formal revenge which was settled on in Chapter 1 was as follows:

RvF. Recipe for Formal Revenge.

We find some semantic notion, λ, constructed in (and, hence, expressible in) the metatheory of $\mathcal{L}_M$ and demonstrate that, were λ expressible in $\mathcal{L}_M$, we could construct a sentence, β, equivalent to $\lambda\langle\beta\rangle \vee \neg T\langle\beta\rangle$, from which we can derive the contradiction $\lambda\langle\beta\rangle \wedge \neg\lambda\langle\beta\rangle$. We demonstrate that this establishes the inexpressibility in $\mathcal{L}_M$ of some notion, σ, which is expressible in $\mathcal{L}$, where it is permitted (indeed, it is common) that $\lambda = \sigma$. We conclude that $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.

So, to get formal revenge on a theory, we construct a semantic notion in the theory's metalanguage and show that it leads, via liar reasoning, to a

contradiction in the object language, and so to the inexpressibility of some notion in the object language which is expressible in natural language. This definition, as discussed in Chapter 2, is deliberately broadened to capture the putative revenge problems which it is argued afflict metatheoretically paraconsistent dialetheism (for example, in Chapter 4 of this thesis). In the case of metatheoretically consistent dialetheism, things are slightly simpler. If we can take a notion expressible in the metalanguage of such a theory and show that, in the object theory, it would lead to contradiction then, since this would generate inconsistency, and so triviality in the classical metalanguage, then that very notion is inexpressible in the object language on pain of triviality. In this particular case, then, the allowance in RvF that the contradiction arising from λ in the object language causes the inexpressibility of σ, where σ may not be identical to λ, is not required. If λ generates inconsistency, and so triviality, if it's expressible in the object language, then λ itself is the inexpressible, on pain of triviality, notion.

Beall argues that metatheoretically consistent dialetheism, or at least his version, does not suffer from revenge. His focus is on the notion 'just true'. I focus on 'just false', though this makes little substantial difference to the discussion. He does not, as I do, distinguish formal from informal revenge. Since the just false problem comes in both varieties, I discuss the formal version of the problem here and the informal version in the next section, and try to point out as I go which of the versions each of his remarks about the problem are aimed at.

Beall's defence of his view draws on his characterisation on revenge given in his (2007) and discussed in Chapter 2 of this thesis. He begins with a discussion of incoherent operators, which he takes to establish what 'just true' (or 'just false') is not. There are certain features, Beall argues, which cannot be had by a notion of 'just true', if it is to be coherent in his object language. Having argued this, he goes on to outlines what he takes 'just true' to be, which is 'true' (similarly, 'just false' is simply 'false').

If we are trying to get formal revenge on a theory, we can construct the revenge notion in the theory's metalanguage in a completely precise, formal way. The features of this notion, then, are not up for grabs in the way that Beall's

discussion of incoherent operators, or the identity of ‘just true’ with ‘true’, seems to assume. I take these remarks, instead, to be best aimed at an *informal* version of the just false problem. So I discuss them in the next section.

Beall begins his discussion of revenge by reminding us that the point of theorising about truth in formal languages is that we wish to provide a model, in the informal sense, of natural language which, we hope, illuminates the features of natural language in which we are interested. One important feature which interests us, for example, is how it is that natural language avoids collapsing into triviality from a combination of liar paradoxes and the apparent correctness of an explosive logic (containing, for example, disjunctive syllogism).

We construct a formal, model language, $\mathcal{L}_M$ characterise a notion of truth in that language, $\mathcal{L}_M$-truth. If the model is a good one, $\mathcal{L}_M$-truth will tell us about $\mathcal{L}$-truth, which is the behaviour of truth in our natural language, $\mathcal{L}$. The revenge problem arises, for metatheoretically consistent dialetheism, from notions which can be constructed in the classical metatheory, like *true-in-*$\mathcal{L}_M$ or *not-true-in-*$\mathcal{L}_M$. Since these notions are completely consistent in the metatheory, we can think of these notions, respectively, as *just-true-in-*$\mathcal{L}_M$ and *just-false-in-*$\mathcal{L}_M$, since they are equivalent.

If these notions were expressible in the object language, we could construct a sentence, λ, equivalent to *just-false-in-*$\mathcal{L}_M$<λ>. The familiar liar reasoning, applied to λ, generates a contradiction, not just in the object theory, but in the metatheory, since, in the metatheory, we will have that λ is both *true-in-*$\mathcal{L}_M$ and *not-true-in-*$\mathcal{L}_M$. This delivers triviality, and so, on pain of triviality, $\mathcal{L}_M$ cannot express *just-false-in-*$\mathcal{L}_M$.

This, I would have thought, would be uncontroversial, even among dialetheists. In fact, Beall questions whether the result goes through (2009, p.56). If, as he supposes is the case, the model language not only has classical set theory as its metalanguage, but contains classical set theory as a proper part, then the revenge argument must fail. The reason is that, if *just-false-in-*$\mathcal{L}_M$ is expressible in the metatheory, it is expressible in classical set theory (since this is what is used for the metatheory), but since the object language contains classical set theory, *just-false-in-*$\mathcal{L}_M$ must be expressible in it. At most, what the argument

establishes is that notions like *true-in-*$\mathcal{L}_M$ do not play the role in $\mathcal{L}_M$ played by $\mathcal{L}_M$'s truth predicate (e.g. transparency).

Exactly where Beall takes the argument to triviality from the expressibility in $\mathcal{L}_M$ of *true-in-*$\mathcal{L}_M$ to break down is slightly unclear. The reference to transparency suggests that it is in the assumption that a sentence being true in the actual model means it's *true-in-*$\mathcal{L}_M$ (and vice versa), though, in fact, we should not conflate this with the notion's being transparent, which is something further. But this is guaranteed by our stipulation that the definition, in the metalanguage, of *true-in-*$\mathcal{L}_M$ is such that *a* is *true-in-*$\mathcal{L}_M$ if and only if α is true at the actual world (@ $\Vdash \alpha$). Beall's point may be that what the argument to triviality shows is not that *true-in-*$\mathcal{L}_M$ fails to be expressible in $\mathcal{L}_M$, but rather that this equivalence cannot hold of it in $\mathcal{L}_M$ in full generality. But, since we define *true-in-*$\mathcal{L}_M$ in terms of this equivalence, then the demonstration that it cannot hold of a notion expressible in $\mathcal{L}_M$ is a demonstration that *true-in-*$\mathcal{L}_M$ can't be expressed in $\mathcal{L}_M$. The fact that some other notion, differently defined, is expressible in $\mathcal{L}_M$ is irrelevant.

So, there are notions expressible in the metalanguage of $\mathcal{L}_M$, and hence expressible in natural language, but which are not expressible in $\mathcal{L}_M$. Beall accepts that this may be so, or at least that the notions in question cannot behave, in the object language, as one might expect. He doesn't think this generates a revenge problem for his theory. His discussion of the issue is similar to the remarks in his (2007) about which putative revenge problems should be taken to be genuine revenge problems. I discussed this article in some detail in Chapter 2 of this thesis, and disagreed with Beall about which problems are genuine revenge problems. In particular, Beall dismissed as 'too easy' putative revenge problems like the one given above for his view. The reasons why this problem is not 'too easy' are the same as in Chapter 2, but I will briefly rehearse them here, in the particular context of Beall's view.

The first issue is that Beall says that the relevance of the inexpressibility result, involving as it does 'classically constructed' notions of the metalanguage, is unclear. For a fuller discussion of this point, I refer the reading to Chapter 2, Section 2.2. Simply put, however, the relevance of the result is that $\mathcal{L}_M$ avoids the threat of triviality only because of certain expressive limitations. There are

notions which, were they expressible, would trivialise the theory and which, therefore, the theory cannot express (assuming it's not trivial). Natural language is not like this, since it can express the notions in question. So $\mathcal{L}_M$ does not tell us how $\mathcal{L}$ (our natural language) avoids the threat of triviality at the hands of the liar paradox, because, demonstrably, it does so differently from $\mathcal{L}$.

One thing which, according to Beall, would turn the objection given above into a genuine, serious revenge problem, is the assumption that the semantics of $\mathcal{L}_M$ are intended to provide a model for the semantics of natural language. Were this assumption made, says Beall, the absence of any notion in $\mathcal{L}_M$ corresponding to *true-in-*$\mathcal{L}_M$, despite the latter notion playing an important role in giving the semantics of $\mathcal{L}_M$, would be a serious problem for $\mathcal{L}_M$ as a model of our natural language, $\mathcal{L}$ (2009, p.56).

Beall, however, rejects this assumption: the semantics of $\mathcal{L}_M$ are *not* intended, on his view, to reflect the semantics of $\mathcal{L}$. The semantics of $\mathcal{L}_M$ are given truth-conditionally in terms of truth-in-a-model. But, as a deflationist, Beall rejects that giving a proper account of the semantics of natural language is a matter of giving truth conditions, since this would require, he thinks, truth to play some more full-blooded, explanatory role than he is willing to accept. The deflationist still faces questions about the logical behaviour of the truth predicate in light of paradoxes like the liar. The deflationist is at liberty, in Beall's view, to answer these questions by describing a language, $\mathcal{L}_M$, truth-conditionally, using model theory. But since the real semantics of $\mathcal{L}$ are not to be understood in this way, the truth conditional semantics are provided in a purely instrumentalist way as giving an account of the logical behaviour of the truth predicate. So the logical behaviour of the truth predicate in $\mathcal{L}_M$ is supposed to model the logical behaviour of the truth predicate in $\mathcal{L}$, but the semantics of $\mathcal{L}_M$ more generally, are not supposed to reflect those of natural language. Beall admits that this instrumentalist account is, at best, promissory, and we still await a model language which does accurately capture the semantics of natural language. Beall is not prepared to give such a theory, but cites the use-theoretic account given by Field (2001) as a promising start. In the meantime, he thinks describing a model language, $\mathcal{L}_M$, truth-conditionally, with the intention of capturing only the logical behaviour of the truth predicate, is innocuous.

As I argued in Chapter 2 of this thesis, especially in Section 2.2, this instrumentalist response to the problem of revenge is, as it stands, inadequate. As Beall admits, we need to be given some model language, say, $\mathcal{L}_U$, whose semantics are given use-theoretically in such a way as to give a model of the semantics of natural language. Moreover, this account of the semantics of natural language cannot be wholly divorced from the features of language intended to be modelled by $\mathcal{L}_M$, such as the veracity of the T-scheme, and the logic which the language obeys. If Beall thinks truth is transparent, and the correct logic is *BXTT*, then we should, if it is to be a model of natural language, be able to describe $\mathcal{L}_U$, with its use-theoretic semantics, in such a way that it has these features too. So the logic of $\mathcal{L}_U$ ought to be *BXTT*, and it should contain a transparent truth predicate (which is, in fact, guaranteed by *BXTT*, it being *BX* with transparent truth). Nothing less than this would satisfy us that Beall's account of the features of the semantics of natural language (given use-theoretically) is compatible with his account of the logical behaviour of the truth predicate.

I see no reason to think, and indeed, it seems extremely unlikely, that $\mathcal{L}_U$, constructed in this way, would avoid the problems $\mathcal{L}_M$ is now facing. That is, it seems very likely that there will be notions, perhaps *not-true-in-*$\mathcal{L}_U$, which would be expressible in $\mathcal{L}_U$'s metalanguage, and hence in natural language, but which would be inexpressible in $\mathcal{L}_U$. So even if we granted that the instrumentalist response got metatheoretically consistent dialetheists off-the-hook as regards revenge for $\mathcal{L}_M$ (though, I have argued, it doesn't), it would have done so only because the theory is problematically incomplete. Beall has given us no reason to think that the completed theory, given by $\mathcal{L}_U$, will avoid the same problem. If it does not, then Beall is committed to this being a genuine revenge problem, since the semantics of $\mathcal{L}_U$, unlike those of $\mathcal{L}_M$, *are* intended to model those of natural language.

So, I have argued, metatheoretically consistent dialetheism is subject to the problem of formal revenge. There are notions expressible in the metalanguage, and thereby expressible in natural language, which cannot be expressed in the object language of the theory. One such notion is *not-true-in-*$\mathcal{L}_M$, which, given the consistency of the metatheory in which it is defined, we can think of as *just-*

false-in-$\mathcal{L}_M$ (since none of the things which are *not-true-in-$\mathcal{L}_M$* will also be *true-in-$\mathcal{L}_M$*). So we can think of the revenge problem given in this chapter as a formal version of the just false problem. As I argue in the next section, metatheoretically consistent dialetheism also faces an informal just false problem. Indeed, this is probably the more familiar way of thinking of the just false problem. I turn to this issue next.

6.2 Informal Revenge for Metatheoretically Consistent Dialetheism

In Chapter 2 of this thesis (Section 2.3), I gave a recipe for informal revenge:

RvI. Recipe for Informal Revenge.

We find some semantic notion, λ, which is not constructed in the metatheory of $\mathcal{L}_M$, but which is expressible in $\mathcal{L}$ and demonstrate that, were λ expressible in $\mathcal{L}_M$, we could construct a sentence, β, equivalent to $\lambda\langle\beta\rangle \vee \neg T\langle\beta\rangle$, from which we can derive the contradiction $\lambda\langle\beta\rangle \wedge \neg\lambda\langle\beta\rangle$. We demonstrate that this establishes the inexpressibility in $\mathcal{L}_M$ of some notion, σ, which is expressible in $\mathcal{L}$, where it is permitted (indeed, it is common) that $\lambda = \sigma$. We conclude $\mathcal{L}_M$ is an inadequate model of $\mathcal{L}$, since it fails to explain how $\mathcal{L}$ has its target features.

As in the case of formal revenge, the consistency of the metatheory, on the sort of view under discussion, makes things slightly simpler than RvI suggests, since the latter was complicated so as to capture revenge problems for metatheoretically paraconsistent dialetheism. In problems of formal revenge, we construct a notion in the metalanguage and show that it cannot be expressed in the object language. Since anything expressible in the metalanguage is expressible in natural language, there are problematic expressive asymmetries between the object language and the natural language it is supposed to model. With informal revenge, we omit appeal to the metalanguage, and appeal directly to the expressibility of a notion in natural language which, via liar reasoning, we can show to be inexpressible in the object language. So the problematic expressive asymmetries are demonstrated without any intermediate appeal to the metalanguage.

In some ways, it is more difficult to press an informal revenge objection against a theory than a formal one. The main reason is that, in the latter case, we can precisely define a notion in the metalanguage and show that the very same notion is inexpressible in the object language. With informal revenge, things are slightly more difficult. What, exactly, the features of a notion found in natural language should be taken to be can, and often are, a matter of dispute. If certain features of a notion lead to it being inexpressible in a dialetheist's object theory, they may simply deny that the notion, properly understood, has those features. If this is implausible, they may simply reject the notion as incoherent. One way to avoid this situation, for those wishing to press the revenge problem, is to look to the dialetheist's own informal remarks about their theory and find there some notion which they cannot properly express in their object language. Especially if the notion appears importantly, and often, the response that it should be discarded as incoherent is likely to seem implausible, and would result in our discarding as incoherent some number of their own remarks about their theory. This is the reason we call this sort of revenge 'informal revenge', because the notion we wish to show is inexpressible is found in the dialetheists informal characterisation of their view. One important notion of this kind is 'just false', as well as related notions like 'just true', 'non-dialetheia' and so on. I focus here, for simplicity, and consistency with the rest of the thesis, on 'just false'. As I have said, Beall focuses, for the most part, on 'just true', but the differences are fairly unimportant in this context.

Beall's discussion of the issue begins with what he takes 'just true' (and 'just false') *not* to be, and with some remarks about incoherent operators, which also draws on some material from his (2007), and which I discussed in Chapter 2 (Section 2.4). Here are some conditions which, according to Beall (2009, p.49), cannot be coherently satisfied by any language:

F1. The language contains a truth predicate, $T\langle x\rangle$, which obeys, unrestrictedly, the T-scheme (at least in its rule form)

F2. Reasoning by cases is valid

E1. $\vDash \alpha \vee \varphi\alpha$

E2. $\alpha, \varphi\alpha \vDash \bot$

Since Beall endorses each of F1, F2 and E1, he must, he says, reject the existence of any operator (at least unrestrictedly) satisfying E2. We might be tempted to introduce some notion, defining 'just true' (or 'just false') such that it is both exclusive (that is, it satisfies E1) and exhaustive (it satisfies E2). But, Beall says, this would lead to triviality via a liar paradox constructed for the notion (and the assumption that F1 and F2 hold). So, Beall says, whatever 'just true' or 'just false' are, they cannot be such as to be both exclusive and exhaustive.

This is a little quick. Firstly, as I have pointed out (2.4), the argument that triviality follows from the above conditions assumes that the standard structural rules are present in the logic. But one might give this up, as do Weir (2013), Ripley (2013) and Zardini (2014). Secondly, it does not follow from the fact that Beall accepts F1 and F2 that 'just true' and 'just false' are neither exclusive nor exhaustive. It may instead be the case that they are, but, because he endorses F1 and F2, plus the standard structural rules, Beall's theory can't express the notions. His discussion presumes the notion to be expressible in his theory, and then subsequently works out, on this assumption, what the notion must be. But whether 'just true' and 'just false' are, for Beall, expressible, is exactly what is at issue.

The obvious understanding of the notion 'just false' is that it be defined as 'false and not true'. Similarly, the obvious understanding of 'just true' is as 'true and not false'. Given Beall's transparent truth predicate, these are equivalent to 'false' and 'true', respectively. So, on Beall's view 'just false' is just 'false' and 'just true' is just 'true'. The problems with this sort of position were discussed in Chapter 4. One obvious way of characterising dialetheism (of the non-trivialist variety) is as the view according to which some things are both true and false, some just true and some just false. But on Beall's account of 'just true' and 'just false', this is equivalent to 'some things are both true and false, some true and some false'. But these sentences are obviously not equivalent. The whole point of uttering the first sentence is to draw a *distinction* between the dialetheia, on the one hand, and the true non-dialetheia, and false non-dialetheia on the other. If Beall's view cannot capture 'just false' so that it

allows for distinctions of this kind, then his view cannot express the notion of sole falsehood.

Beall seems to accept that there must be *some* distinction between a sentence's being false and its being just false and invokes a pragmatic device to try to capture it. If one says that a sentence is just false, then this is exactly equivalent so saying that it is false, but carries the pragmatic implicature that one also *rejects* the sentence. For Beall, then, the 'just' in 'just false' carries an 'autobiographical implicature' which the mere utterance of 'false' does not (2009, p.52).

This response does not solve the just false problem for the reasons given by Shapiro in his (2004). The notion 'just false' can be embedded in conditionals, or into hypotheses and negated, but pragmatic devices cannot (or at least, we have no account of how they could). I may be considering endorsing dialetheism and, in my deliberations about the view, wish to entertain the hypothesis that it is just false. How I could do this without Beall's 'autobiographical' account of the notion is unclear.

A further difficulty is that the pragmatic implicature is tied to the speakers own rejection behaviour[9]. On the face of it, the notion 'just false' is one of semantic classification. There are some sentences, on dialetheism, which fall into the 'dialetheia' category, some others fall into the 'just true' category, and some further sentences which fall into the 'just false' category. But Beall's suggested pragmatic device is not like this: it appeals to the subjective mental attitudes of the speaker, not the objective semantic features of the theory itself. If 'just false' is an important classificatory notion employed in the articulation of the theory, as it seems to be, and one thinks a theory ought to be, in some sense, detachable from its proponents, then one might think this problematic. It is an unattractive feature of a theory of the paradoxes that an articulation of it makes essential reference to the mental states of its proponents. One might try to avoid this by making the rejection involved normative. Perhaps, one might think, the pragmatic device points to the fact that the 'just false' sentence *ought* to

[9] One might wish to make a distinction between acceptance/rejection, on the one hand, and assertion/denial, on the other. Acceptance and rejection are mental attitudes one can take towards some content. Assertion and denial are the speech acts which express them. The differences, however, are not crucial in this context.

be rejected. The problem with this is that it immediately invites further paradoxes: for example, the sentence, *β*, equivalent to '*β* ought to be rejected'. Assuming a few plausible principles governing correct rejection, this sentence ought to be both rejected and accepted.

Priest has accepted this sort of rational dilemma (2006, p.274), and a metatheoretically consistent dialetheist is equally entitled to this view. In the present circumstances, however, this would have the effect of making 'just false' behave inconsistently again, defeating the purpose of introducing the pragmatic device in the first place.

A final attempt Beall makes to assuage potential worries about the equivalence of 'false' with 'just false' appeals, again, to his deflationism. The truth predicate was not introduced to name an important feature of the world, but as a logical tool to overcome our expressive limitations. Before recognising dialetheia, Beall says, we took it to be obvious that 'false' and 'just false' are equivalent, and after recognising dialetheia, the equivalence remains. According to Beall, if we thought that truth was some genuine property of things in the world, one might expect notions like 'just true' or 'just false' to be similar and so to be distinct from true or false, respectively. Since he is a deflationist, he does not think this, and so takes 'just true' to be 'true' and 'just false' to be 'false'.

It doesn't seem to me that these considerations help much. It may be that those who do not accept dialetheia think it obvious that 'false' and 'just false' (defined as 'false and not true') are equivalent. On the other hand, since they don't accept dialetheia, they have no need of the notion of sole falsity. It was only when discussion of dialetheism began that the notion became worthy of discussion, largely because dialetheists must make use of it in characterising their view. If dialetheia are not accepted, there is no plausible sense in which there is a distinction between being false and being just false, so it is to be expected that no one attempts to draw such a distinction. Once dialetheia are accepted (and so long as one doesn't accept trivialism), there is a very plausible sense in which being false and being just false and distinct. So appealing to the fact that, without dialetheia, there is no distinction between falsity and sole falsity, gives no support whatever to the claim that, once dialetheia are

recognised, there should be no distinction. It is precisely the introduction of dialetheia which makes sense of, and requires, a distinction between ‘false’ and ‘just false’.

The appeal to deflationism does not seem to help either. Some sentences, for Beall, are both true and false, some are just true and some are just false. This claim captures a central claim of his view and, to make sense of it, there must be distinction between being false and being just false (or being true and being just true). This is so no matter how deflated the notion of truth involved in the claim is taken to be. So, it seems to me, the invocation of deflationism here is a red herring.

The metatheoretically consistent dialetheist may attempt to avoid this result by denying one of the principles which lead to the equivalency between ‘false’ and ‘just false’, perhaps the *exclusion* principle that $F\langle\alpha\rangle \rightarrow \neg T\langle\alpha\rangle$, which is rejected by Priest, and discussed in Chapter 4. Matters are much the same in the context of metatheoretically inconsistent dialetheism, so I refer the reader there for more details. The upshot of the discussion is that, if one rejects exclusion, one can keep ‘false’ and ‘just false’ distinct, though the latter is still inconsistent (infinitely many sentences which are just false are also true). The problems with this approach are that it is still unclear whether the resulting notion still, despite its inconsistency, counts as a genuine notion of sole falsity, and that one is forced to give up the transparency of the T-scheme and the plausible principle that falsity entails untruth. Other problems, too, are mentioned in the sections cited.

Beall’s settled view is that a proper treatment of sole falsity should be as equivalent to falsity. I have argued that this is unsatisfactory. He does, however, offer two further suggestions available to merely semantic dialetheism, and they are equally available to any metatheoretically consistent dialetheist (2009, pp.58-62). He admits that the suggestions are ‘speculative’ and it seems to me that they are inadequate in fairly straightforward ways, so I discuss them only briefly.

The first involves a hierarchy of partial ‘just false’ predicates. The first, which we might call JF_0 applies to sentences of the (truth-free) base language. We then

add a further predicate, JF_1, which we can predicate of the sentences in the fragment containing the base language and any sentences obtained by substituting $T\langle\beta\rangle$ for any β in α where α and β are sentences of the base language. We call this fragment $\mathcal{L}_1$ and construct further languages (or fragments), $\mathcal{L}_i$, and corresponding 'just false' predicates JF_i, for each i, in a similar way. The effect is a hierarchy of 'just false' predicates, each applied to an increasingly large fragment of the language. There are two problems with this approach. The first is that invoking hierarchies of semantic predicates decreases the attractiveness of dialetheism, part of the promise of which is an advance on hierarchical theories. If we are going to invoke a hierarchy of 'just false' (as well as 'just true') predicates into the language, one might think, we might as well introduce a hierarchy of truth predicates and endorse Tarski's view, without the bother of dialetheism. The second is, as with all hierarchical views, we immediately encounter revenge problems. We can express in natural language, for example, the idea of a sentence being just false at every level, i, in the hierarchy. This can't be expressed in Beall's hierarchy. So this response to the just false problem simply moves the problem elsewhere.

Beall second suggestion is to introduce a notion of 'negation' which has only a partial definition. In particular, it has only a sufficient condition for its application. I don't discuss the details of the suggested operator here, since the problem with the notion doesn't depend on them (they can be found in his (2009, p.59-62)). Setting aside the theoretical unattractiveness of such partially defined operators, the central problem with the suggestion is that the notion of negation he employs is inconsistent. That is, letting '#' stand for Beall's 'negation', there are sentence, such as some liar sentence, α, equivalent to $\#T\langle\alpha\rangle$ such that both α and $\#T\langle\alpha\rangle$ hold. So, invoking # to define 'just false', does not prevent the inconsistency of the latter: there will still be infinitely many sentences which are both just false and true.

One option a dialetheist like Beall might wish to consider is simply rejecting the notion of just false as incoherent. Beall suggests this in his discussion of potential revenge objections in his (2007). He says that whoever wishes to demonstrate that a view is subject to revenge must demonstrate that the inexpressible notion really is an intelligible notion of natural language, and that

it may be perfectly reasonable for those subject to the potential problem to simply deny this (2007, p.13).

Unfortunately for Beall, this option is not available. He often uses expressions like ‘just true’ and ‘just false’ and even more frequent use of notions which seem to depend on notions like ‘just false’, ‘just true’ or ‘non-dialetheia’ for their meaning. To pick some arbitrary examples, he says:

“Rational dialetheists maintain that some (actually, many) [truths] are *just true*; they reject that all or even most claims are [dialetheia]. Indeed, on my account, it is only the spandrels of [truth] (or related notions) that are [dialetheia]; the rest are ‘just true’ (2009, 48).

He says this in the introduction to his chapter addressing the just false (or, just true) problem. The notion of sole truth is explicitly mentioned twice, and his claim that it is *only* the spandrels of truth which are dialetheia would seem to rely on the notion as well since, presumably, we should understand the ‘only’ to mean that the other sentences are *non-dialetheia*.

Elsewhere, he says that, on his view “it’s just the [truth]-ineliminable spandrels that wind up [as dialetheia]” (2009, p.24). He also says “the [dialetheia] – the [true falsehoods] – are essentially tied to our given see-through device [truth] (or related notions).” (2009, p.6) Similarly, he claims, when giving the formal semantics for his view that “the [dialetheia] in the theory arise only in (the weird parts) of the semantic fragment.” (2009, p.13) There are many other such examples in his writing, as there are in the writing of all dialetheists.

Though these last three examples do not explicitly use the expression ‘just false’, ‘just true’ or ‘non-dialetheia’, it’s difficult to see how they could be understood without them. After all, what does it mean to say that it is *just* the spandrels that are dialetheia, or that dialetheia are *essentially* tied to truth, or that dialetheia *only* arise in the semantic fragment, if not that the spandrels are dialetheia, but the non-spandrel, non-semantic sentences which are not essentially tied to truth are *non-dialetheia*?

Given the pervasive use of notions of this kind in the literature on dialetheism, it is simply not plausible for a dialetheist to claim that the notions in question are

incoherent. Dialetheists of the sensible variety think that, though there are true contradictions, these are a rarity: most sentences we encounter in the course of a normal day are *not like this*. So long as this is the case, it behoves dialetheists to say things about these other, normal sentences, and this requires notions which categorise them like ‘just true’, ‘just false’ and ‘non-dialetheia’.

6.2.1 Shrieking Just False

In some recent papers JC Beall has articulated what he calls ‘The Shrieking Method’, employing so-called ‘Shriek Rules’ to do various jobs for dialetheists. In two of these, he uses Shriek Rules to obtain ‘classical recapture’ results ((2013a), (2013b)). In a third, which is my concern here, he attempts to use Shriek Rules to solve the ‘just true’ and ‘just false’ problems (2013c). I argue in this section that shriek rules do not help the dialetheist avoid the ‘just false’ problem. One thing which will help make this clear is a more careful account of the distinct-but-related problems of ‘just false’ and ‘exclusion’. These are often conflated in the literature, and they are importantly conflated in Beall’s paper.

The just false problem is, essentially, one of semantic categorisation. There are lots of sentences which are false and not, in addition, true but, it is alleged, dialetheists have no notion which captures these sentences. One simply way to capture dialetheism of the more sensible, non-trivialist variety, is as the view according to which some sentences are both true and false, some are just true and some are just false. Dialetheists ought to be able to express this claim and, focusing on the part concerning ‘just false’, they ought to be able to express claims like

(*JF*) Some sentences are false, but some are just false.

As I have pointed out, sentences of this kind are extremely common in discussions of dialetheism, and are uttered frequently both by dialetheists are their opponents. The just false problem arises when we consider how ‘just false’ behaves, given its obvious definition as ‘false but not true’, on dialetheism. Assuming, as Beall does, that a sentence is false if and only if it is untrue, ‘just false’ ends up equivalent to false and, for all the distinction the notion draws, the dialetheist might as well have said

(*JF**) Some sentences are false, but some are false.

JF and *JF** are obviously inequivalent: the second misses something crucial about the first, which is that the 'just' adds something important to the first sentence. Any view which treats *JF* as completely equivalent to *JF** cannot express the notion 'just false'.

If one doesn't take falsity to be equivalent to untruth and, in particular, if one denies that a sentence's being false entails its untruth (that is, if one denies *exclusion*), then things are slightly different. On this view, falsity and just falsity are inequivalent notions, since a sentence can be false without being untrue and, therefore, without being just false. However, if the converse, that a sentence's being untrue entails that it is false, holds, then any untruth is thereby false and, hence, just false. So any sentence which is untrue, including a great many (indeed, infinitely many) dialetheia, is also just false. Though dropping exclusion restores the inequivalency between *JF* and *JF**, it's not clear that this solves the problem. If one rereads *JF* with the understanding that 'just false', and it occurs there, applies to infinitely many true sentences, it's not implausible to think that something has still been missed. What we wanted was to say something about the sentences which are *just* false, not the dialetheic ones which are also true.

As discussed in Chapter 4 of this thesis, one response which is available here, and the one endorsed by Graham Priest, is to deny that the inconsistency of the notion prevents it from meaning 'just false'. I think this is wrong, but as I said in the cited section, it is difficult to present arguments to this effect beginning with premises the dialetheist is likely to accept. It seems to me that it is built-in to the meaning of 'just false' that it behave consistently. Priest thinks otherwise. Perhaps some empirical work, which checked speakers' patterns of assent/dissent to various uses of the term in various contexts, might make progress on the issue. But this work has not been carried out, and, in the meantime, it seems likely that dialetheists will simply deny that it is part of the meaning of the term that is behave consistently, leaving us with an, at least temporary, stalemate.

On the other hand, dialetheists are in the business of trying to convert other philosophers to their view, and, without having done a careful survey, I suspect that most philosophers' intuitions will side with mine. This being so, it would be dialectically advantageous for dialetheist to have a consistent notion of 'just false', should they want it. It's difficult to see how this could be a bad thing, at any rate. So there is some reason for dialetheists to try to characterise a consistent notion of just false.

However, Priest's response to the problem, to accept the inconsistency of just falsity, is only available to him because he denies the exclusion principle. This has some unattractive features, one of which being that one cannot have transparent truth, as Beall does. If one accepts that a sentence is false if and only if it is untrue, this collapses the distinction between falsity and just falsity, and, therefore, the distinction between *JF* and *JF**. Such views, then, cannot, if 'just false' is defined as above, express sole falsity.

So the 'just false' problem is that dialetheists, on the face of it, cannot have a notion of semantic categorisation which respects the difference between sentences like *JF* and sentences like *JF**.

While the 'just false' problem is concerns dialetheism's ability to classify sentences by semantic value, the exclusion problem concerns their ability, as the name suggests, to *exclude* or *rule-out* certain sentences. Exactly what this should amount to is not always ideally clear, but one way of getting at the problem is through considerations of disagreement. In his (2004, p.338), Shapiro argues (citing Parsons (1990) as the original source of the objection) that dialetheists cannot express disagreement. Suppose I am a dialetheist, and someone makes a claim, α, with which I wish to disagree. I might try to express my disagreement by saying '$\neg\alpha$' but, though $\neg\alpha$ is inconsistent with α, it is not, by my dialetheist lights, *incompatible* with α. But I cannot disagree with something by asserting something else with which it is compatible. What I need, one might think, is some way of *excluding* or *ruling-out* the truth of α, in such a way that what I say is genuinely incompatible with α. Only if this can be done, the thought goes, can I express my disagreement with α. The challenge, then, is for the dialetheist to find some device which allows us to exclude, and which

behaves consistently, so that when I use the device to exclude α this is not compatible with α.

One response on behalf of dialetheism, made by Graham Priest, is to point out that there is a sense in which, say, a classicist is no better off:

"A dialetheist can express the claim that something, α, is not true – in those very words, $\neg T\langle\alpha\rangle$. What she cannot do is ensure that the words she utters behave consistently: even if $\neg T\langle\alpha\rangle$ holds, $\alpha \wedge \neg T\langle\alpha\rangle$ may yet hold. But in fact, a classical logician can do no better. He can endorse $\neg T\langle\alpha\rangle$, but this does not prevent his endorsing α as well. . . . [C]lassical logic, as such, is no guard against this. . . [A]ll the classical logician can do by way of saying something to indicate that α is not to be accepted is to assert something that will collapse things into triviality if he does accept α. But the dialetheist can do this too. She can assert $\alpha \rightarrow \bot$." (2006: 291)

Suppose, now, that I am a classicist, and I wish to disagree with an assertion that α. I do so by asserting that $\neg\alpha$. According to Priest, it's not clear that I've really *ruled-out* α, since I might yet accept both α and $\neg\alpha$. It's a consequence of the explosiveness of classical logic that this would commit me to triviality, but classical logic, one might think, is not guard against this *per se*.

To take an example, Paul Kabay's (2010) book 'On the Plenitude of Truth' is a defence of trivialism, on the basis that dialetheism is true, and classical logic is correct. So whilst trivialism strikes most as, to say the least, theoretically unpalatable, it doesn't look like *classical logic itself* precludes it.

Nonetheless, perhaps the fact that classical logic delivers triviality in the presence of contradictions is, for most, a form of exclusion. If I, as a classicist, respond to someone's assertion that α with 'it is not the case that α', then to the extent that trivialism is theoretically 'off the table' (as it is for everyone apart from Paul Kabay, so far as I am aware), α is 'off the table' too (at least for me). In Beall's words, I've ruled-out α 'up to triviality' (2013c, p.440).

Ruling-out ‘up to triviality’ is something that classicists get for free, since explosion is valid in the logic they endorse. Dialetheists don’t get it for free (assuming they’re of the more conservative, paraconsistent variety), at least in most cases. If a dialetheist accepted ‘Everything is true’, even they would be committed to triviality. So perhaps if a dialetheist were to say ‘It’s not the case that everything is true’, we would take them to have ruled-out ‘Everything is true’, up to triviality. But for the most part, this will not be possible.

Priest’s suggestion is that this feature can simply be added to dialetheism by saying of the thing, α, we wish to exclude, $\alpha \rightarrow \bot$. In later work, Priest rejects this strategy, and I discuss the reasons for this presently. He revises his view to invoke a *sui generis* speech act of denial. This, it seems to me, is a slightly different sort of exclusion to that which rules out ‘up to triviality’, so I’ll give a rough distinction between two senses in which we might wish to exclude. The first, I’ll call ‘disagreement exclusion’. As the name suggests, this is the sort of exclusion one wishes to use purely for the purposes of expressing conversational disagreement of the kind mentioned above: someone asserts α and if one disagrees, one wants a speech act of some kind which allows one to express this disagreement. This is the sort of disagreement for which one might invoke a speech act like denial. But some philosophers want something more than this. For example, Franz Berto (2014) characterises a notion of primitive, metaphysical exclusion which holds between properties. Two properties can exclude one another irrespective of whether any speaker utters a sentence expressing the fact, and independent of any person’s mental attitudes towards the properties. I call this sort of exclusion ‘theoretical exclusion’, since it allows us to have a theory which gives an account of what excludes what, whether or not this is put to work by speakers actually expressing disagreement. Beall’s account, which I will characterise shortly, is also, plausibly, of this kind, since it calls for the addition to a theory of non-logical shriek rules characterising what excludes what ‘up to triviality’, on that theory.

Having some notion of theoretical exclusion, plausibly, gives one a notion of disagreement exclusion, since one can simply assert a sentence including the theoretical exclusion device to express disagreement with the sentence, α, with which we wish to disagree. The converse is not obviously true. One might have a notion of disagreement exclusion, such as denial, but not think that this builds-

in to one's theory an account of which things are excluded, independent of individual speakers' actual acts of denial.

For each of the two forms of exclusion, either disagreement or theoretical, there are corresponding exclusion problems to the effect that a dialetheist cannot express that form of exclusion. Though, again, it may be that any reasonable solution to the theoretical exclusion problem also provides a solution to the disagreement exclusion problem (the converse is less plausible). One way to put the difference between the problems is that a theorist faces the *disagreement* exclusion problem if there is no speech act available to them with which they can exclude for the purposes expressing disagreement. On the other hand, a theorist faces the *theoretical* exclusion problem if there is no notion in their theory which, independently of any particular speech acts, gives an account of what is ruled-out by the theory.

These exclusion problems are importantly distinct, however, from the just false problem. Again, the just false problem is one of semantic categorisation. Dialetheists, it is alleged, cannot adequately categorise sentences by semantic value: they frequently, as they must, categorise sentences as 'just false', 'just true', 'non-dialetheia' and so on, but they lack an account of how these notions are expressible on their view. The exclusion problems, however, are different. On these objections, it is argued that dialetheists cannot adequately express that certain things are *excluded* or *ruled-out*.

To take an example, suppose, as Beall does, that exclusion is a matter of ruling-out 'up to triviality'. Consider someone who endorses a paraconsistent logic, but not dialetheism. Perhaps they are motivated by the thought that relevance is important to logical consequence, but not convinced by a dialetheist account of the paradoxes. There is no reason, it seems to me, to charge such a person with the 'just false' problem, because their view is consistent. They do, however, face the theoretical exclusion problem, since it is not a feature of their logic that contradictions entail triviality. So they, just as much as dialetheists, (at least *prima facie*) lack the ability to rule out 'up to triviality'.

I don't wish to rule-out in advance, by mere stipulation, that someone might use the same machinery to address each of these problems and so treat the just false problem and the exclusion problems, essentially, the same. This may well

be possible. On the other hand, since it is certainly not guaranteed that a solution to one will deliver a solution to the others, it's important, at least for the purposes of discussion, to keep the just false problem separate from issues of exclusion.

A final consideration is exactly which things we should expect, in either sense, a dialetheist to be able to exclude. One suggestion, which seems sensible enough, is that they ought to be able to exclude all those things which are just false; this is one point of contact between the just false problem and exclusion.

6.2.1 Priest on Just False and Exclusion

Priest does not conflate the just false problem with exclusion. In the passage quoted above, he says that dialetheists can express that something, α, is not true (by which he means, as is clear from the context, 'just false') using those very words '$\neg T\langle\alpha\rangle$'. This, as is discussed in Chapter 4 of this thesis (Section 5.1.1), is Priest's response to the 'just false' problem: the dialetheist simply uses the expression 'just false' (or, equivalently, for Priest, 'not true') and accepts that the notion behaves inconsistently. He has a separate response to the exclusion problem, which is that we can rule-out 'up to triviality', α, by asserting $\alpha \rightarrow \bot$ (which I will call the 'arrow-falsum' strategy). Plausibly, this is a sort of theoretical exclusion: for any α such that $\alpha \rightarrow \bot$ holds, α is ruled-out 'up to triviality' by the theory. We can then express disagreement exclusion, as Priest says, for any such α by simply asserting $\alpha \rightarrow \bot$.

It's even clearer in another passage from Priest that he treats the problems separately:

"[M]any people have argued that a dialetheic solution to the paradoxes can be maintained only by expressive incompleteness, particularly with respect to the notion of being *false only*…[Dialetheists] cannot assert that something is false-only if this is required to exclude things that are true as well. For both of us [Field and Priest], though, there is a way of obtaining this effect with a different kind of speech act, namely *denial*: both of us can deny that *A* is true. In some contexts there are things that can be asserted which have the same force as a

denial. Thus, both of us can assert $A \rightarrow \bot$, in the face of which one can maintain A only on pain of triviality, which would normally be taken as a denial. As Field points out, though, this cannot be used to reject in all cases. The Curry sentence, *K*, for example, is itself of the form $K \rightarrow \bot$, and so cannot be asserted by way of rejecting *K*. With all of this I agree. There is nothing that can be asserted which will, in general, have the same force as
denial. The speech act is sui generis.

So far, our lines are parallel. However, on further projection, they diverge. The crucial difference concerns what can be said in propositional contexts. Field has literally no way of expressing the notion of indeterminacy…The dialetheist about the paradoxes does have a way of expressing that something is false only – in the very words 'false and not true'. It is just that these cannot be guaranteed to behave consistently." (2010a, pp.136-137)

So Priest endorses, again, the view that the correct solution to the 'just false' problem is to define it in the usual way, and accept its (in Priest's case, more limited) inconsistency. He clearly has a different solution in mind for exclusion. In certain cases, we can exclude by asserting $\alpha \rightarrow \bot$, but more generally, we exclude using denial.

The reason Priest cites for abandoning the arrow-falsum strategy is that there are certain sentences which cannot be excluded in this way. The example he gives, due to Field (2008, pp.388-390) is of a Curry sentence, *K*, equivalent to $K \rightarrow \bot$. This sentence must be false only, and so must be excludable but, if the only available way of excluding is arrow-falsum, then the only way to exclude *K* is to assert $K \rightarrow \bot$. But if this is asserted, then triviality follows by the substitutivity of equivalents and *modus ponens*. So there are some sentences which cannot be excluded using arrow-falsum.

In fact, things are slightly worse than this, as Beall points out (2013c, pp.440-441). The most significant problem with the arrow-falsum strategy, is the strength of '→' which, for Priest, (as is discussed in Chapter 3, Section 3.2.0) is a *logical strength* connective, requiring for $\alpha \rightarrow \beta$ that β follow logically from α. In the context of the arrow-falsum strategy, this means that, for $\alpha \rightarrow \bot$ to be

true, α must logically entail triviality. In other words, it must be the case that there is no non-trivial world at which α holds. To take an example, it may be just false that I am in the park, and so, in that case, a dialetheist ought to be able to exclude my being in the park. But unless my aversion to the park is of a very extreme kind, it won't be the case that my being in the park entails that everything is true. So, on Priest's account of the conditional, since it will almost never be the case that the thing we wish to exclude would, if true, entail triviality, the arrow-falsum strategy can almost never be used to exclude.

As Beall says, one way to respond to this situation would be to weaken the conditional to make it more easily satisfiable. The difficulty with this strategy is that it is easy to weaken the conditional too much. For example, if we interpreted the conditional materially, and so, to rule out α we assert $\alpha \supset \bot$, then, since this is equivalent to $\neg\alpha \vee \bot$, it will be true of every α such that $\neg\alpha$ holds that α is excluded. This would have the effect of 'excluding' every dialetheia there is, which is obviously unsatisfactory. So this strategy would face the task of weakening the conditional such that $\alpha \rightarrow \bot$ holds for all and only those sentences which are just false. This may be possible, but it will not be easy. According to Beall, a simpler, and better, response is to take the core idea of the arrow-falsum strategy, and implement it using 'shriek rules' instead.

6.2.2 Shrieking, Just False and Exclusion

In Beall's notation, we let 'α!' (read 'α shriek') be '$\alpha \wedge \neg\alpha$' and then let $\alpha! \vDash \bot$ be α's shriek rule. I use '$\vDash$' rather than Beall's '$\vdash$', here, to make clear that semantic entailment is involved (as Beall spells out in the more precise, formal accounts of shrieking in his (2013a) and (2013b)). I also, in what follows, drop the 'shriek' notation, and simply use '$\alpha \wedge \neg\alpha \vDash \bot$' as the shriek rule for α, since this seems to me simpler. Though I focus here, for simplicity of presentation, on the sort of simple shriek rule just given, Beall focuses mainly on more complicated shriek rules for predicates, with a view to 'shrieking' whole

theories. The details of these more complicated rules are not required for the points I wish to make, and would distract unnecessarily from them. My focus on simple shriek rules for single sentences is innocuous for two reasons: firstly, we need to be able to say of sentences that they are just false, and to be able to exclude them. If the shrieking strategy cannot be applied to sentences so as to achieve this, it fails to solve either the exclusion problem or the 'just false' problem. Secondly, since theories are simply sets of sentences, we can take single sentences to be perfectly legitimate limiting cases.

Shriek rules are, according to Beall, non-logical rules, which are added to our theory in much the same way an epistemologist might add a factivity principle, $K\varphi \vDash \varphi$, to a formal theory of knowledge. They are added to our theory for all those things we take to be consistent. The effect of adding a shriek rule for α, $\alpha \wedge \neg\alpha \vDash \bot$, is to enforce consistency 'up to triviality' on α. If we wish to rule-out α up to triviality, then once the shriek rule has been added, we can simply assert '$\neg\alpha$'. The assertion of $\neg\alpha$, combined with the shriek rule, has ruled-out α 'up to triviality', since, were α to hold too, triviality would follow.

As should be clear from the above discussion (in which I distinguished the 'just false' problem from issues of exclusion), that this helps at all with the 'just false' problem is not obvious. Recall that the challenge of the just false problem was to find some notion which allowed us to express 'just false' in such a way as to capture the difference between sentences like those I called '*JF*' and '*JF**', respectively:

(*JF*) Some sentences are false, but some are just false.

(*JF**) Some sentences are false, but some are false.

But, on Beall's shrieking strategy, we add shriek rules to our theory, forcing consistency 'up to triviality' on those things we take to be consistent, then, we can assert, of some α we take to be ruled-out, '$\neg\alpha$', or, equivalently, '$F\langle\alpha\rangle$' (or $\neg T\langle\alpha\rangle$, or $T\langle\neg\alpha\rangle$). So shriek rules do nothing to distinguish *JF* from *JF**: they are both, on Beall's shrieking strategy, exactly equivalent. So Beall's claim that shriek rules solve the 'just false' problem seems to rest on the conflation of that problem with the problem of exclusion.

One might respond to this that Beall's view can be modified to address the 'just false' problem, most obviously by taking the 'just' in sentences like *JF* to be somehow indicating the existence of a shriek rule. There are two ways one might try to implement this suggestion. The first is the simplest: we just insert the shriek rule into our utterance and express 'α is just false' by '$F\langle\alpha\rangle \wedge (\alpha \wedge \neg\alpha \vDash \bot)$'. If the sentence expressing the shriek rule '$\alpha \wedge \neg\alpha \vDash \bot$' is, as one would assume, a sentence of the metatheory, this does not help. We already have a consistent notion of sole falsity in the metatheory, and this does not solve the problem. We must, instead, be able to express 'just false' in the object language. To make this point clear, we can introduce a validity predicate, *Val*, into the object language and express shriek rules using this. So α's shriek rule will be $Val(\alpha \wedge \neg\alpha, \bot)$. We can then express, in the object language, that α is just false by '$F\langle\alpha\rangle \wedge Val(\alpha \wedge \neg\alpha, \bot)$'.

This, one might think, would allow Beall's shriek rules to express 'just false', as well as to exclude 'up to triviality'. Unfortunately, introducing the shriek rule into our utterance invites revenge paradox. Consider the sentence, β, equivalent to $F\langle\beta\rangle \wedge Val(\beta \wedge \neg\beta, \bot)$. This sentence must be just false since, if it is true (or both true and false), we can derive triviality as follows:

(1) $T\langle\beta\rangle$	(assumption)
(2) β	(1, T-scheme)
(3) $F\langle\beta\rangle \wedge Val(\beta \wedge \neg\beta, \bot)$	(2, substituting identicals)
(4) $F\langle\beta\rangle$	(3, $\wedge$-elimination)
(5) $T\langle\neg\beta\rangle$	(4, F-scheme)
(6) $\neg\beta$	(5, T-scheme)
(7) $Val(\beta \wedge \neg\beta, \bot)$	(3, $\wedge$-elimination)
(8) $\beta \wedge \neg\beta$	(2, 6, $\wedge$-introduction)
(9) $\bot$	(7, 8, *Val*-detachment)

So *β* must be just false, but, on the present way of expressing this, we assert *F*<*β*> ∧ *Val*(*β* ∧ ¬*β*, ⊥) which, by substituting identicals, applying the T-scheme and ∧-introduction, this gives us *T*<*β*> ∧ (*F*<*β*> ∧ *Val*(*β* ∧ ¬*β*, ⊥)). In other words, *β* is both just false and true. So, if we attempt to express sole falsity by adding a shriek rule to our utterances, the notion ends up inconsistent, defeating the purpose of introducing shriek rules in the first place. This, of course, is a revenge problem for this strategy.

The second strategy makes the familiar appeal to a pragmatic implicature. So, although, strictly speaking the content of *JF* is the same as that of *JF**, the 'just' in the former generates a pragmatic difference, by pointing to the existence of a shriek rule.

This suggestion fails for the, also familiar, reason that it is expressively impoverished: one cannot embed pragmatic implicatures in, for example, hypotheses or conditionals. A second worry about this strategy is that we can assert shriek rules; this being so, we can assert that a sentence is false and that a shriek rule holds of it. If the falsity of a sentence, in combination with a shriek rule, is what makes something 'just false', why should we be forced to invoke a pragmatic device, rather than simply having an expression whose content expresses 'just false' directly? This, presumably, would be preferable. If the only reason for invoking the pragmatic device is to avoid the revenge paradox, involving *β*, which I have just sketched, then it is *ad hoc*.

So Beall's shriek rules do not help with the 'just false' problem. Indeed, they are not aimed at that problem, but at the exclusion problem. If we attempt to modify Beall's shrieking account to address the just false problem, then, again, we are faced with a dilemma: either we have an expression whose content contains the notion 'just false', in which case, revenge paradox delivers its inconsistency; or, we introduce a pragmatic device which is both lacking in motivation and expressively limited.

This still leaves the exclusion problem, at which shriek rules seem better aimed. Unfortunately for Beall, shriek rules do not address this problem either, and for

the same reason that Priest's arrow-falsum strategy fails: they are too strong. On Priest's arrow-falsum strategy, to exclude some sentence, α, we assert $\alpha \rightarrow \bot$. The problem with this, given the strength of Priest's '$\rightarrow$' is that it amounts to the claim that there is no non-trivial world at which α holds. But this is almost never true. All we need to find is a sentence which is false only, and so which requires excluding, but which is true in some other, non-trivial situation.

Beall's shriek rule for some *a* to be excluded, which says $\alpha \wedge \neg\alpha \vDash \bot$, does not require that there be no non-trivial situation at which α holds, but it does require that there be no non-trivial world at which $\alpha \wedge \neg\alpha$ holds. So all we need to do is find a sentence which is just false, but which is both true and false in some non-trivial situation. This will be something which Beall ought to be able to exclude, but which he can't using shriek rules.

Where we find such sentences depends on the variety of dialetheism in question. Some dialetheists, like Graham Priest, think there are dialetheia in lots of places: they arise from the law, from puzzles concerning vagueness, from puzzles concerning the borders of objects, the paradoxes of change and so on. It seems likely that all the dialetheia which are supposed to arise in such cases are going to be contingent. For example, suppose I am well-inside a room with a hall outside. At the moment, it is just false that I am in the hall. But, if Priest is right about borders, then at the instant at which my location precisely lies on the boundary between the room and the hall, I am both in the hall and not in the hall. This situation, of course (assuming, again that Priest is right), is also non-trivial. So although it is just false that I am in the hall, and so Beall ought to be able to exclude my being in the hall, he cannot do so using shriek rules, for there are non-trivial situations in which I am both in the hall and not.

I suspect that if we endorse any metaphysical dialetheia whatsoever (arising in the law, from vague predicates, etc.), we're going to end up with sentences which are just false, but which could have been both true and false (non-trivially), and so with counter-examples to the shrieking strategy, since these are things we ought to be able to exclude, but which we can't shriek.

Beall, of course, does not accept the existence of such dialetheia, restricting his inconsistency to the semantic realm. But this doesn't prevent the problem, since there is a well-known class of contingent semantic paradoxes: the empirical liars. These are sentences which have the potential to become genuine liar sentences and so, for Beall, genuine dialetheia, but which require some other, empirical facts to obtain for this to be the case. If these facts do not obtain, the sentences may be just true, or just false. Some examples of such sentences are as follows:

a) (*) The sentence marked (*) is false
b) The third sentence from the bottom of the page is false
c) There is a sentence on the page which is false
d) Either this sentence is false, or it's raining
e) Everything Nixon said is true

Depending purely on contingent, empirical matters (like which other sentences are on the page, how they're marked, what their truth-values are, whether it's raining, what Nixon said), these sentences might be dialetheia, or they might be true only or false only. If they're false only, say, we ought to be able to exclude them (if they're true only, we might want to exclude their negation). But we can't do so on the shrieking strategy, because there are non-trivial cases such that they both hold and don't.

So Beall's shrieking strategy, like Priest's arrow-falsum strategy, fails because it is too strong. There are sentences we ought to be able to exclude, but which we cannot. Importantly, these counterexamples count just as much against shriek rules as a solution to the 'just false' problem as they do against them as a solution to exclusion. So even if Beall was able to negotiate the dilemma given above for the shrieking response to 'just false', it would still fail, because these counterexamples are just false, but can't be expressed as such using shriek rules.

6.3 Conclusion

This chapter has argued that versions of dialetheism which take a consistent metatheory suffer from the problem of revenge. My particular focus has been on Beall's theory, offered in his *Spandrels of Truth*, as the most thorough articulation of the view. The revenge problems his theory faces are more general, however. By taking a classical metatheory, the dialetheist gives rise to notions expressible in that metatheory which cannot be expressed in the object theory. The object theory is, therefore, an inadequate model of natural language, since the former avoids triviality differently from the latter. This is what I have called 'formal revenge'. Beall has argued that this sort of revenge is 'too easy' but, as I have argued, this is not the case. The theory also faces informal revenge at the hands of the 'just false' problem. Dialetheists must treat this notion inconsistently, and attempts to remedy the situation by introducing a consistent notion of sole falsity face a dilemma: either the notion can be contained in an expression, in which case revenge paradox delivers its inconsistency; or appeal is made to a pragmatic device, in which case we are back with expressive limitation.

Though, I have argued, metatheoretically paraconsistent dialetheism also faces the problem of revenge, that theory has some *prima facie* advantages over its metatheoretically consistent counterpart. I sketched a few of these in above. Perhaps the most important of these is that metatheoretically consistent forms of dialetheism are denied Priest's arguments for naïve set theory; for example, that a proper understanding of absolutely general quantification requires a universal set. These arguments are powerful, and, if dialetheism cannot, as I have argued, be motivated by appeal to its immunity from the problem of revenge, these seem the most promising avenue for an alternative motivation for the view. It is a cost to metatheoretically consistent dialetheism that it cannot endorse these.

Bibliography

Emil Badici (2008). The liar paradox and the inclosure schema. *Australasian Journal of Philosophy* 86 (4):583 – 596.

Diderik Batens (1990). Against global paraconsistency. *Studies in Soviet Thought* 39 (3-4):209-229.

J. C. Beall (2009). *Spandrels of Truth*. Oxford University Press.

J. C. Beall (2007). Prolegomenon to future revenge. In , *Revenge of the Liar: New Essays on the Paradox*. Oxford University Press.

Jc Beall (2013a). Free of Detachment: Logic, Rationality, and Gluts. *Noûs* 48 (3).

Jc Beall (2013b). Lp+, k3+, fde+, and their 'classical collapse'. *Review of Symbolic Logic* 6 (4):742-754.

Jc Beall (2013c). Shrieking against gluts: the solution to the 'just true' problem. *Analysis* 73 (3):438-445.

Jc Beall (2011). Dialetheists against Pinocchio. *Analysis* 71 (4):689-691.

Jc Beall , Thomas Forster & Jeremy Seligman (2013). A Note on Freedom from Detachment in the Logic of Paradox. *Notre Dame Journal of Formal Logic* 54 (1):15-20.

Jc Beall & Julien Murzi (2011). Two Flavors of Curry Paradox. *Journal of Philosophy* 110 (3):143-165.

Jc Beall & Greg Restall (2006). *Logical Pluralism*. Oxford University Press.
Francesco Berto (2014). Absolute Contradiction, Dialetheism, and Revenge. *Review of Symbolic Logic* 7 (2):193-207.

Francesco Berto & Graham Priest (2008). Dialetheism. *Stanford Encyclopedia of Philosophy* (2008).

George Boolos (1998). Reply to Charles Parsons' ``sets and classes". In Richard Jeffrey (ed.), *Logic, Logic, and Logic*. Harvard University Press. 30-36.

Ross T. Brady (1989). The non-triviality of dialectical set theory. In G. Priest, R. Routley & J. Norman (eds.), *Paraconsistent Logic: Essays on the Inconsistent*. Philosophia Verlag. 437--470.

Lewis Carroll (1895). What the tortoise said to Achilles. *Mind* 4 (14):278-280.

Jon Cogburn (2004). The Philosophical Basis of What? The Anti-Realist Route to Dialetheism. In Graham Priest, J. C. Beall & Bradley Armour-Garb (eds.), *The Law of Non-Contradiction*. Clarendon Press.

Donald Davidson (1967). Truth and meaning. *Synthese* 17 (1):304-323.

K. Devlin (1980). *Fundamentals of Contemporary Set Theory*. Springer-Verlag

Michael Dummett (1973). *Frege: Philosophy of Language*. Duckworth.

Michael A. E. Dummett (1978). *Truth and Other Enigmas*. Harvard University Press.

P. Eldridge-Smith (2012). Pinocchio beards the Barber. *Analysis* 72 (4):749-752.

P. Eldridge-Smith (2011). Pinocchio against the dialetheists. *Analysis* 71 (2):306-308.

Hartry Field (2001). *Truth and the Absence of Fact*. Oxford University Press.

Hartry H. Field (2008). *Saving Truth From Paradox*. Oxford University Press.

Jay Garfield (2004). To Pee and not to Pee? Could That Be the Question? (Further Reflections of the Dog). In Graham Priest, J. C. Beall & Bradley Armour-Garb (eds.), *The Law of Non-Contradiction*. Clarendon Press.

I. Grattan-Guinness (1998). Discussion. Structural similarity of structuralism? Comments on Priest's analysis of the paradoxes of self-reference. *Mind* 107 (428):823-834.

Anil Gupta (1982). Truth and paradox. *Journal of Philosophical Logic* 11 (1):1-60.
Richard Heck (2007). Self-reference and the languages of arithmetic. *Philosophia Mathematica* 15 (1):1-29.

Hans G. Herzberger (1982). Notes on naive semantics. *Journal of Philosophical Logic* 11 (1):61 - 102.

Paul Horwich (2005). *Reflections on Meaning*. Oxford University Press,Clarendon Press ;.

Paul Kabay (2010). *On the Plenitude of Truth: A Defense of Trivialism*. LAP Lambert Academic Publishing

Saul A. Kripke (1975). Outline of a theory of truth. *Journal of Philosophy* 72 (19):690-716.

Frederick Kroon (2004). Realism and Dialetheism. In Graham Priest, J. C. Beall & Bradley Armour-Garb (eds.), *The Law of Non-Contradiction*. Clarendon Press.

David Lewis (1991). *Parts of Classes*. Blackwell.

David Lewis (1970). General semantics. *Synthese* 22 (1-2):18--67.

Greg Littman & Keith Simmons (2004). A critique of dialetheism. In G. Priest, J. C. Beall & B. Armour-Garb (eds.), *The Law of Non-Contradiction*. Oxford University Press.

Edwin D. Mares (2004). Semantic Dialetheism. In Graham Priest, J. C. Beall & Bradley Armour-Garb (eds.), *The Law of Non-Contradiction*. Clarendon Press.

John Mayberry (1977). On the consistency problem for set theory: An essay on the Cantorian foundations of classical mathematics (I). *British Journal for the Philosophy of Science* 28 (1):1-34.

Alexius Meinong (1960). On the theory of objects (translation of `Über Gegenstandstheorie', 1904). In Roderick Chisholm (ed.), *Realism and the Background of Phenomenology*. Free Press. 76-117.

Richard Montague (1974). *Formal Philosophy; Selected Papers of Richard Montague*. New Haven,Yale University Press.

Chris Mortensen (2013). Motion perception as inconsistent. *Philosophical Psychology* 26 (6):913-924.

Chris Mortensen (1997). The Leibniz continuity condition, inconsistency and quantum dynamics. *Journal of Philosophical Logic* 26 (4):377-389.

Chris Mortensen (1985). The limits of change. *Australasian Journal of Philosophy* 63 (1):1 – 10.

Terence Parsons (1990). True Contradictions. *Canadian Journal of Philosophy* 20 (3):335 - 353.

Graham Priest (2014). *One: Being an Investigation into the Unity of Reality and of its Parts, including the Singular Object which is Nothingness*. Oxford University Press

Graham Priest (2013). Indefinite Extensibility—Dialetheic Style. *Studia Logica* 101 (6):1263-1275.

Graham Priest (2012). Definition Inclosed: A Reply to Zhong. *Australasian Journal of Philosophy* 90 (4):789 - 795.

Graham Priest (2010a). Hopes fade for saving truth. *Philosophy* 85 (1):109-140.

Graham Priest (2010b). Inclosures, Vagueness, and Self-Reference. *Notre Dame Journal of Formal Logic* 51 (1):69-84.

Graham Priest (2008). Logical pluralism hollandaise. *Australasian Journal of Logic* 6:210-214.

Graham Priest (2006). *Doubt Truth to Be a Liar*. Oxford University Press.

Graham Priest (2006). *In Contradiction: A Study of the Transconsistent*. Oxford University Press.

Graham Priest (2005). *Towards Non-Being: The Logic and Metaphysics of Intentionality*. Oxford University Press.

Graham Priest (2002). *Beyond the Limits of Thought*. Oxford University Press.

Graham Priest (2001). *Introduction to Non-Classical Logic*. Cambridge University Press.

Graham Priest (2000). On the principle of uniform solution: A reply to Smith. *Mind* 109 (433):123-126.

Graham Priest (1998). The import of inclosure: Some comments on Grattan-guinness. *Mind* 107 (428):835-840.

Graham Priest (1994). The structure of the paradoxes of self-reference. *Mind* 103 (409):25-34.

Graham Priest (1987) 'Unstable Solutions to the Liar Paradox' in *Self Reference: Reflections and Reflexivity*, S.J. Bartlett and P. Suber (eds.), Nijhoff

Graham Priest (1979). Logic of Paradox. *Journal of Philosophical Logic* 8 (1):219-241.

F. P. Ramsey (1990). *F.P. Ramsey: Philosophical Papers*. Cambridge: Cambridge University Press.

Agustín Rayo & Gabriel Uzquiano (eds.) (2006). *Absolute Generality*. Oxford University Press.

Stephen Read (2006). Monism: The One True Logic. In D. de Vidi & T. Kenyon (eds.), *A Logical Approach to Philosophy: Essays in Memory of Graham Solomon*. Springer.

Greg Restall (1997). Paraconsistent logics! *Bulletin of the Section of Logic* 26 (3):156-163.

Greg Restall (1992). A note on naive set theory in ${rm LP}$. *Notre Dame Journal of Formal Logic* 33 (3):422-432.

David Ripley (2013). Paradoxes and Failures of Cut. *Australasian Journal of Philosophy* 91 (1):139 - 164.

David Ripley (2011). Negation, Denial, and Rejection. *Philosophy Compass* 6 (9):622-629.

R. Routley & V. Routley (1972). The semantics of first degree entailment. *Noûs* 6 (4):335-359.

Richard Routley , Val Plumwood , Robert K. Meyer & Ross T. Brady (1982). *Relevant Logics and Their Rivals*. Ridgeview.

Bertrand Russell (1986). *The Philosophy of Logical Atomism and Other Essays, 1914-19*. Allen & Unwin.

Kevin Scharp (2013). *Replacing Truth*. Oxford University Press.

Stewart Shapiro (2007). Burali-Forti's revenge. In J. C. Beall (ed.), *Revenge of the Liar: New Essays on the Paradox*. Oxford University Press.

Stewart Shapiro (2004). Simple truth, contradiction, and consistency. In G. Priest, J. C. Beall & B. Armour-Garb (eds.), *The Law of Non-Contradiction*. Oxford University Press.

B. H. Slater (1995). Paraconsistent logics? *Journal of Philosophical Logic* 24 (4):451 - 454.

Nicholas J. J. Smith (2000). The principle of uniform solution (of the paradoxes of self-reference). *Mind* 109 (433):117-122.

Richard Sylvan (1980). *Exploring Meinong's Jungle and Beyond: An Investigation of Noneism and the Theory of Items*. Research School of Social Sciences, Australian National University.

Alfred Tarski (1936). The concept of truth in formalized languages. In A. Tarski (ed.), *Logic, Semantics, Metamathematics*. Oxford University Press. 152--278.

Zach Weber (2013). Notes on inconsistent set theory. In. In Francesco Berto, Edwin Mares, Koji Tanaka & Francesco Paoli (eds.), *Paraconsistency: Logic and Applications*. Springer. 315--328.

Zach Weber (2012). Transfinite cardinals in paraconsistent set theory. *Review of Symbolic Logic* 5 (2):269-293.

Zach Weber (2010a). Transfinite numbers in paraconsistent set theory. *Review of Symbolic Logic* 3 (1):71-92.

Zach Weber (2010b). Extensionality and restriction in naive set theory. *Studia Logica* 94 (1):87 - 104.

Zach Weber (2010c). Explanation and Solution in the Inclosure Argument. *Australasian Journal of Philosophy* 88 (2):353-357

A. Weir (1998). Naïve set theory is innocent! *Mind* 107 (428):763-798.

Alan Weir (2013). A Robust Non-transitive Logic. *Topoi*:1-9.

Alan Weir (2004). There Are No True Contradictions. In Graham Priest, J. C. Beall & Bradley Armour-Garb (eds.), *The Law of Non-Contradiction*. Clarendon Press.

Elia Zardini (2014). Naive truth and naive logical properties. *Review of Symbolic Logic* 7 (2):351-384.

Haixia Zhong (2012). Definability and the Structure of Logical Paradoxes. *Australasian Journal of Philosophy* 90 (4):779 - 788.