9

Revenge, Field, and ZF

Graham Priest

9.1 Introduction

This chapter deals with three interrelated issues:

1. What is the 'revenge' phenomenon?
2. How does it bear upon Field's account of the semantic paradoxes?
3. Is the notion applicable to the set theoretic paradoxes?

The meanings of these questions, and the connections between them, will become clear in due course.

9.2 Revenge

Let us start with a statement of the simple liar paradox. This concerns a statment, L, of the form $F\langle L\rangle$, where F is the falsity predicate, and angle brackets indicate some name-forming device. If T is the truth predicate, then the T-scheme assures us that for any closed sentence, A:

$$T\langle A\rangle \leftrightarrow A$$

Substituting L, we have:

$$T\langle L\rangle \leftrightarrow F\langle L\rangle$$

Then applying the Principle of Bivalence for L, $T\langle L\rangle \vee F\langle L\rangle$, we infer $T\langle L\rangle \wedge F\langle L\rangle$, which contradicts the Principle of Univalence applied to L, $\neg(T\langle L\rangle \wedge F\langle L\rangle)$.

When people attempt to give an account of the paradoxes of semantic self-reference, such as the liar, they invoke certain machinery (a theory of truth, truth-value gaps, revision, etc.). It would seem to be that in many, if not all, cases this machinery can be deployed to formulate a related version of the paradox, just as virulent as the original. This is what I shall understand, for the purpose of this chapter at least, as the 'revenge' phenomenon.

There is, in fact, a uniform method for constructing the revenge paradox—or extended paradox, as it is called sometimes. All semantic accounts have a bunch of Good Guys (the true, the stably true, the ultimately true, or whatever). These are the ones that we target when we assert. Then there's the Rest. The extended liar is a sentence, produced by some diagonalizing construction, which says of itself just that it's in the Rest. The diagonal construction, because of its ability to tear through any consistent boundary, may then play havoc. This shows, incidentally, that the extended paradox is not really a different paradox. The pristine liar is the result of the construction when the theoretical framework is the standard one (all sentences are true or false, not both, and not neither). 'Extended paradoxes' are simply the results of applying the construction in different theoretical frameworks.

To see what options there are for handling the revenge situation, it is useful to look at it from the following perspective. The semantic paradoxes arise, in the first instance, as arguments couched in natural language. One who would solve the paradoxes must show that the semantic paradoxes do not, despite appearances, lead to contradiction (or, at least, triviality in the case of a dialetheic approach, but let us focus on consistent approaches for the moment; I will comment on the dialetheic case later). And it is necessary to show this for every concept in the semantic family, for they are all deeply implicated in paradox. Attempts to do this, given the resources of modern logic, all show how, given a language, $\mathcal{L}$, to construct a theory, $\mathcal{T}_{\mathcal{L}}$, for the semantic notions of $\mathcal{L}$, according to which they behave consistently. We now have a series of options.

Horn 1 Are the concepts of $\mathcal{T}_{\mathcal{L}}$ expressible in $\mathcal{L}$? If the answer to this is 'yes', it always seems possible to use the resources of $\mathcal{T}_{\mathcal{L}}$ to formulate the extended paradox, and so obtain a contradiction. Neither is this an accident. For since the concepts of $\mathcal{T}_{\mathcal{L}}$ are expressible in $\mathcal{T}_{\mathcal{L}}$, and since, according to $\mathcal{T}_{\mathcal{L}}$, things are consistent, we should be able to prove the consistency of $\mathcal{T}_{\mathcal{L}}$ in $\mathcal{T}_{\mathcal{L}}$. And provided that $\mathcal{T}_{\mathcal{L}}$ is strong enough in other ways—for example, if it contains the resources of arithmetic—then we know that $\mathcal{T}_{\mathcal{L}}$ is liable to be inconsistent, by Gödel's second incompleteness theorem. The upshot of this case is, then, inconsistency.

Horn 2 If the answer to our original question is 'no', the concepts of $\mathcal{T}_{\mathcal{L}}$ are not expressible in $\mathcal{L}$. In this case, we ask another question: are they expressible in some

other language? If the answer is yes, then $\mathcal{L}$ is expressively incomplete. There are certain semantic concepts that it cannot express. But then, in that case, the original problem of showing that our semantic concepts behave themselves has not been solved. For $\mathcal{T}_{\mathcal{L}}$ deals only with the semantic concepts of $\mathcal{L}$, and now there are others, also prone to generate inconsistency—as horn 1 of the situation shows—that have not been dealt with.

Horn 3 The other possible answer to this question is 'no': the concepts in question are not expressible at all. If this is to be a robust theoretical position, and not to lapse into gesturing at the ineffible, we must insist that the concepts in question are not meaningful; they do not exist. This is the case of inexistence (non-existence). At first blush, this position would seem to be shamelessly self-refuting, since the theorist has depended upon those very notions in giving their own account. But things are not quite so straightforward. We have talked simply of the concepts of $\mathcal{T}_{\mathcal{L}}$. But there are two ways in which a theory can invoke concepts. It may do so explicitly, by giving them names, reasoning about them, etc. If the semantic concepts of $\mathcal{L}$ are invoked in this way, then we do indeed have immediate self-refutation. But, more subtly, the concepts may not be invoked explicitly: they may be presupposed in some way, thus being invoked implicitly. If this is the case, denying the meaningfulness of the concepts is an option, if one can sustain the view that the concepts are not *really* presupposed. How problematic this move is now turns on how robust the presupposition is. If it is one that is hard to gainsay, the theory would appear to be in just as much trouble.

Just to illustrate this last possibility, think of Wittgenstein's *Tractatus*. The statements of the *Tractatus* say that a certain kind of sentence is meaningless (*unsinnig*). According to the *Tractatus*, most of the statements of the *Tractatus* are of this kind. But the *Tractatus*, though it does not *say* this, would seem to presuppose that these sentences *are* meaningful. It uses them. In the end, though, Wittgenstein simply denies the presupposition, and insists that they really are meaningless. I do not think that the move can be sustained coherently. But we need not go into that now. The situation illustrates how an account may presuppose something that it does not explicitly assert, how that presupposition may be denied, and the problems that this sort of move is wont to give.[1]

The upshot of the preceding discussion is this. The revenge scenario poses the theorist with three possibilities: inconsistency, incompleteness, and inexistence, each with its own characteristic problems. One or other of these horns must be selected and coped with.

[1] The Tractarian situation is certainly one of self-reference, though it is not a standard paradox of self-reference. Its structure is, however, one of an inclosure; so it is the same form of the paradoxes of semantic self-reference. For further discussion, see Priest (1995), ch. 12.

9.3 An Illustration

This is all rather abstract. Let me illustrate it with an example.[2] Consider the Tarskian solution to the semantic paradoxes. We start with a language, $\mathcal{L}_0$, with no semantic concepts. To reason about the semantics of $\mathcal{L}_0$, we move to a different language, $\mathcal{L}_1$, which contains the truth predicate, T_0, applying to the sentences of $\mathcal{L}_0$. That is, we have all instances of the T -schema, $T_0 \langle A \rangle \leftrightarrow A$, where A is a sentence of $\mathcal{L}_0$. To reason about the semantics of $\mathcal{L}_1$, we have to repeat the move, generating a hierarchy of language, which may be depicted as follows:

Language	T-Schema	Legitimate Instances
⋮	⋮	⋮
$\mathcal{L}_{i+1}$	$T_i \langle A \rangle \leftrightarrow A$	$A \in \mathcal{L}_i$
⋮	⋮	⋮
$\mathcal{L}_2$	$T_1 \langle A \rangle \leftrightarrow A$	$A \in \mathcal{L}_1$
$\mathcal{L}_1$	$T_0 \langle A \rangle \leftrightarrow A$	$A \in \mathcal{L}_0$
$\mathcal{L}_0$	None	

Given techniques of self-reference, it is easy enough to construct a sentence, L, such that $L = \neg T_i \langle L \rangle$. If we had $T_i \langle L \rangle \leftrightarrow \neg T_i \langle L \rangle$, and given that we have the Law of Excluded Middle, $A \vee \neg A$, we would have a contradiction. But we do not. L is a sentence of $\mathcal{L}_{i+1}$; so we have only $T_{i+1} \langle L \rangle \leftrightarrow \neg T_i \langle L \rangle$, and the liar-reasoning is blocked. More generally, given an appropriate formal definition of the hierarchy, and an interpretation for $\mathcal{L}_0$, one can prove that the hierarchy of truth theories is consistent.

Call the level in the hierarchy at which a sentence appears (or the first such level if the levels are cumulative, as they are usually taken to be) its *rank*. Let $rk(x)$ be the rank of x. The Good Guys in this construction are the ones that are true at their rank; that is, the sentences that satisfy the predicate $T_{rk(x)}x$. So the extended liar is one that says of itself that it is not true at its rank:

$$L: \neg T_{rk(\langle L \rangle)} \langle L \rangle$$

We now have the possibilities corresponding to the three horns.

The first is that the notion of being true at its rank is expressible in some language in the hierarchy. Suppose this is $\mathcal{L}_i$. Then L is a sentence of $\mathcal{L}_i$, and $rk(\langle L \rangle) = i$. So at level $i + 1$, we have $T_i \langle L \rangle \leftrightarrow \neg T_{rk(\langle L \rangle)} \langle L \rangle \leftrightarrow \neg T_i \langle L \rangle$, and we have contradiction. This is the inconsistency case.

[2] For further examples, see Priest (1987), ch. 2, and the 2nd edn., 19.3.

The second is that the predicate $T_{rk(x)}x$, though meaningful, cannot be expressed in the hierarchy. What this shows is that there are semantic concepts with the potential to generate contradiction, and which are not dealt with in the theory. This is the incompleteness case.

The third is the inexistence case. We can deny that there is such a concept as 'truth at its rank', that it is a meaningful notion. In the Tarskian construction, the concepts employed in its expression are invoked explicitly by the theorist; hence a denial of its existence (meaningfulness) is a simple self-refutation. (To specify the hierarchy, the theorist must say that for each level of the hierarchy, y, there is a truth predicate, T_y, at level y. So quantification into the subscript place of the truth predicates must be legitimate.)

Before we move on to Field, let me conclude with a word on dialetheism and revenge. In a dialetheic treatment of the semantic paradoxes, the Good Guys are the truths. The Rest are the things that are false but not also true (assuming for the sake of argument that there are no truth-value gaps). So the extended liar is a sentence, L, of the form $F\langle L\rangle \wedge \neg T\langle L\rangle$. Assuming all these concepts to be expressible in the language, reasoning about this sentence in the natural way one can demonstrate that $F\langle L\rangle \wedge T\langle L\rangle \wedge \neg T\langle L\rangle$.[3] This is a contradiction. So the revenge phenomenon applies just as much to the dialetheic theory. We end up in Horn 1, inconsistency. But clearly, though this horn of the dilemma is devastating for consistent accounts of the paradoxes, it is not so for dialetheic ones.

9.4 Field

With this background, let us now move to Field's account of the semantic paradoxes, and see what happens there. I will not give an exegesis of his account. It can be found in Field (2003), (2005), and (2007).

Since the question of what can and what cannot be expressed in a language is clearly crucial, let us start by getting clear what the language of Field, the theorist, is. It is essentially the language of *ZF* augmented by a truth predicate, T, and a non-truth-functional conditional, $\rightarrow$, to be deployed in stating the T-schema. Field gives a semantics for this language. According to the interpretations he defines, the purely set-theoretic vocabulary behaves classically; we may therefore reason as in classical *ZF* as long as T and $\rightarrow$ are not being used (as opposed to mentioned) in the reasoning. Generally, though, the semantics are many-valued, with the unique designated value, 1, which is the value of the Good Guys (the things that we can

[3] See Priest (1987), 2nd edn., 20.3.

correctly assert). The semantics can all be described in *ZF*, however, so it is permissible to reason classically about it.

In this context, the extended paradox is clearly generated by a sentence, *L*, of the form $Val(\langle L \rangle) \neq 1$, where $Val(x)$ is 'the value of x', and having value 1 marks out the robustly acceptable sentences of the language. Now we have our three familiar possibilities: inconsistency, incompleteness, and inexistence.

Field's preferred option (expressed most explicitly in discussion) is, in fact, Horn 3, inexistence. The notion makes no sense. At first blush, this looks like a straightforward self-refutation. Hasn't Field shown how to define the predicate in set-theoretic terms? No. What he has done is shown how, given any interpretation of *ZF*, $\mathcal{M}$, to define the predicate $Val_{\mathcal{M}}(x) = 1$, 'x has value 1 relative to interpretation $\mathcal{M}$'. It is the absolute notion, which is not explicitly defined in the construction (nor can it be, or we would have a consistency proof for the theory, and so for *ZF*, in *ZF*), the existence of which Field denies.

This notion would appear to be presupposed by the construction in a very robust way, though. The whole point of Field's construction is to delineate and justify the inferences we are allowed to deploy in the language in question (Field's own language). It will do so only if the language has a semantic structure of the same kind as that of the interpretations that Field specifies. As someone (I forget who) said, truth in a model must be a model of truth. Having value 1 in a model must be a model of having value 1. If there is no such notion, then the fact that $Val_{\mathcal{M}}$ can be deployed to give a certain notion of validity provides *no reason whatsoever* to suppose that the notion applies to Field's language. Clearly, Field thinks it does, since he often appeals to the notion of validity he delineates to justify the legitimacy or otherwise of forms of reasoning in the language he uses. If he denies the existence of this notion, he cannot claim that certain ways of reasoning in the language he uses are legitimate—or illegitimate. Since he does make such claims, and takes his semantics to justify these, he does presuppose the notion.

There is more to be said on this matter, but before we turn to this, let us look briefly at the other two horns, inconsistency and incompleteness. The first of these is that *Val* is expressible in the language. If it is expressible in the purely set-theoretic language, then we have, for any A, $Val(\langle A \rangle) = 1 \vee Val(\langle A \rangle) \neq 1$, and the familiar classical reasoning gives a contradiction. If *Val* is definable in the language, but not the purely set-theoretic language, then we need not have all instances of the Law of Excluded Middle for $Val(\langle A \rangle) = 1$, and the argument to paradox is blocked. But we may now legitimately ask how to define it. We cannot define it using Field's D operator. Even within an interpretation, $\mathcal{M}$, is not the case that this applies to all and only the things that have value 1 in $\mathcal{M}$ (and the same goes for all the other operators in Field's D hierarchy). But even given a definition, we have a dilemma. If $Val(\langle A \rangle) = 1 \vee Val(\langle A \rangle) \neq 1$, we have the paradox with us; if not, our previous

worry is exacerbated. For any $\mathcal{M}$, $Val_{\mathcal{M}}(\langle A \rangle) = 1 \vee Val_{\mathcal{M}}(\langle A \rangle) \neq 1$; so having value 1 in a model is *not* a good model of having value 1.

The final possibility is the incomplete case: *Val* is a meaningful notion, but not expressible in the language. In that case, the solution fails for the standard reason: the construction has not shown potentially inconsistency-generating semantic notions to be inconsistent. It should be noted that the Definition of *Val* can be carried out in second-order *ZF*—not the second-order version of Field's theory; just second-order *ZF*. We can simply apply Field's construction to the model of first-order *ZF* that second-order *ZF* gives us.[4] Hence, Field must deny the legitimacy of second-order *ZF*, a point which considerably ratchets up the stakes in the inexistence horn.

One way or another, then, Field is subject to the revenge syndrome.[5]

9.5 Revenge and ZF

The problem with Field's preferred approach, as we have seen, is that he denies the existence of a notion that, in all honesty, he has to presuppose. Note, however, that this is not a problem of Field's theory as such. It is simply one that the theory inherits from orthodox set-theory, *ZF*.

Suppose, *per*, one might hope, *impossibile*, that one could define the intended interpretation of the language of *ZF* in *ZF*. Then Field's construction would give us an intended interpretation for his extended language, and his theory of validity would be applicable to its own language. The failure to be able to do this is, then, simply a result of a certain inability of *ZF*.

Since *ZF* is normally taken to be perfectly kosher, this might suggest that there is nothing to worry about. But there is. If one were to attempt to specify the intended interpretation for *ZF* in *ZF*, it would have the structure $\langle V, \in_V \rangle$, where V is the set of all sets (or, assuming the Axiom of Foundation, the Cumulative Hierarchy), and $\in_V$ is the membership relation on V. But one cannot do this, since V is not a set. Now, what are we to say of this V? There are three possibilities.

1. Inconsistency. The first is that the existence of V can be recognized in *ZF*. In some way, we can prove its existence. In this case, of course, *ZF* would be inconsistent. We would be able to prove the consistency of *ZF* in *ZF*, and so Gödel's second incompleteness theorem would kick in. This is like using a sledge hammer to crack a nut, however. Since *ZF* entails the non-existence of V, we have an immediate contradiction.

[4] See Rayo (2007).

[5] This is not the only problem with Field's account. For a discussion of revenge and other problems, see Priest (2006).

2. *Incompleteness.* The second is to suppose that V exists, but that it is not one of the sets in *ZF*. But this would show that *ZF* is not the theory of *all* collections, which it was supposed to be. Nor does it help to suppose that V is a proper class, assuming this notion to make sense. Proper classes would seem to be just the next layer up in the cumulative hierarchy. If the sets of *ZF* really exhaust the hierarchy, there is no next layer. Even if there were, exactly the same problem would then arise with respect to the totality of all classes (proper and otherwise). So the problem has not been solved; merely relocated.

3. *Inexistence.* For that reason, the only really robust possibility for a solution is to deny the existence of V *tout court*. But that's problematic. Reasoning in natural ways, we would appear to make legitimate use of large totalities such as V on many occasions. The tendency to invoke proper classes is but a manifestation of this fact. For example, the natural understanding of various categories in category theory, such as the category of all sets (let alone the category of all categories) is exactly about such totalities.

And *ZF* itself would appear to *presuppose* the existence of V. For a start, the model-theoretic account of validity given in *ZF* cannot be applied to the language of *ZF* itself unless $\langle V, \in_V \rangle$ is an interpretation, which it is not. In other words, just as for Field, the logic the theory defines is not applicable to the theory itself, and we are bereft of a justification for reasoning about sets, one way or the other.[6]

Other considerations point in the same direction. A quantified sentence has no determinate truth-value unless the range of the quantifiers is a determinate totality. If I say 'everyone has the right to vote', what I say is true if restricted to adults, false if it includes minors. But in *ZF* we quantify over all sets, and we take it that the theorems of *ZF* are determinately true. There must, then, be a determinate totality of all sets. (Call this a proper class if you want. The name is unimportant.) Just as in the case of the *Tractatus*, denying the existence of V is therefore tantamount to denying that statements of *ZF* have meaning—or at least determinate meaning. For a second reason, then, *ZF* seems to presuppose a totality the existence of which it denies.[7]

[6] The problem is well recognized by classical logicians. One solution is proposed by Kreisel (1967). By appealing to a pre-theoretic notion of validity and its supposed properties, he argues that we may take the absolute notion of validity to be extensionally equivalent to the model-theoretic notion. One might have various objections to Kreisel's argument; but in any case, the strategy is unlikely to appeal to Field, just because he, unlike Kreisel, is trying to drive a wedge between the model-theoretic situation and the absolute situation. In particular, the absolute notion of validity, presupposing as it does the function *Val*, must also, according to him, be meaningless.

[7] Further on all these matters, see Priest (1987), ch. 2, and (1997), ch. 11. The problem discussed in this section is essentially that generated by Cantor's paradox of the greatest cardinal size. But, as one might expect, Burali–Forti's paradox of the greatest ordinal size produces essentially the same problem. See Shapiro (2007).

The revenge problem is normally thought of as applying to the semantic paradoxes, not the set-theoretic paradoxes, but as we have just seen, the revenge situation is exactly the same for set theory, at least for *ZF*.[8] The set V is not itself a semantic notion, but it is conceptually closely connected with the semantics of *ZF*.[9] And one has the same three options: inconsistency, incompleteness, and inexistence—each with its own come-uppance.

Any solution to the semantic paradoxes which piggybacks on *ZF*, such as Field's, whatever it says about truth, is, therefore, subject to revenge problems. The revenge of V.

References

Beall, JC, and Armour-Garb, B. (2006). *Deflationism and Paradox*, Oxford: Oxford University Press

Field, H. (2003). 'A revenge-immune solution to the semantic paradoxes', *Journal of Philosophical Logic* 32, 139–77

_____ (2006). 'Is the liar both true and false?', ch. 2 of Beall and Armour Garb (2006)

_____ (2007). 'Solving the paradoxes, escaping revenge', this volume

Kreisel, G. (1967). 'Informal rigour and completeness proofs', pp. 138–71 of I. Lakatos (ed.), *Problems in the Philosophy of Mathematics*, Amsterdam: North Holland

Priest, G. (1987). *In Contradiction*, Dordrecht: Martinus Nijhoff. 2nd edn. Oxford: Oxford University Press (2006)

_____ (1995). *Beyond the Limits of Thought*, Cambridge: Cambridge University Press. 2nd edn. Oxford: Oxford University Press (2002)

_____ (2006). 'Spiking the Field artillery', ch. 3 of Beall and Armour Garb (2003)

Rayo, A. and Welch, P. D. (2007). 'Field on revenge', this volume

Shapiro, S. (2007). 'Burali–Forti's revenge', this volume

[8] There are some non-standard set theories, such as Quine's *NF*, which have a universal set. Obviously, the considerations I have applied to *ZF* do not carry over immediately to them. However, there will be similar considerations in such cases. For example, although such theories may well be able to define their own intended interpretation, what, then, they cannot do, on pain of inconsistency, is prove that the theory holds in that interpretation. Again, we are bereft of a justification for supposing that the theory applies.

[9] See Priest (1987), 2.5.