Prolegomenon to future revenge

JC Beall

University of Connecticut

jc.beall@uconn.edu

Final Draft | January 29, 2007 | Revenge of the Liar | Oxford: OUP

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

1 Truth

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

CAPTURE: $\alpha \Rightarrow Tr(\ulcorner\alpha\urcorner)$

RELEASE: $Tr(\ulcorner\alpha\urcorner) \Rightarrow \alpha$

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

The names 'Capture' and 'Release' arise from the fact that $Tr(x)$ *captures* the information in x, fully storing it for its eventual *release*. In practice, a familiar—if not *the*—role of $Tr(x)$ is its release function: an assertion of $Tr(\ulcorner\alpha\urcorner)$ releases all of the information in α. This is useful for 'long generalizations' or 'blind generalizations' or the like, many of which generalizations would be practically impossible if we didn't enjoy a truth predicate that played

(at least the rule form of) Capture and Release.[1] That truth plays Capture and Release in Conditional Form is plausible but controversial. That truth plays Capture and Release in at least Rule Form is less controversial, and will henceforth be assumed.[2]

2 The Liar

The Liar phenomenon involves sentences that imply their own falsity or, more generally, untruth. By way of example, consider the ticked sentence in §2 of this essay.

√ The ticked sentence in §2 of 'Prolegomenon to future revenge' is not true.

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

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

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

3 The Liar's Revenge

On one hand, the *revenge phenomenon*—the Liar's revenge—is not so much a distinct phenomenon from the Liar as it is a witness to both the difficulty and ubiquity of Liars. On the other hand, 'revenge' is often launched as an *objection* to an account of truth (or a response to Liars). Without intending a stark distinction, I will discuss *revenge qua Liar phenomenon* and *revenge qua objection* separately, with most of the discussion on the latter but all of the discussion brief.

3.1 The Revenge phenomenon

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

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

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

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

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

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

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

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

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

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

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

⋆ The starred sentence in §3.1 of 'Prolegomenon to future revenge' is either not true or a bugger.

Assume that the starred sentence is true. Release delivers that the starred sentence is not true or a bugger, and hence true *and* either not true or a bugger. Similarly, that the starred sentence is either not true or a bugger implies, via Capture, that it is true—and, hence, true *and* either not true or a bugger. Accordingly, given normal

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

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

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

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

A *quietist* advises that we give up on our aim to classify Liars; there are Liars in the language, but there is no 'semantic category' in which the ticked sentence may truly be said to reside. Accordingly, whereof one cannot truly classify, thereof one must—or, in any event, might as well—be silent. The virtue of such an approach is that it avoids revenge (since it doesn't engage); the salient defect is that it offers no clear account of truth or the paradoxes at all. Against such a 'proposal', little can be said, and so won't.

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

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

3.2 Towards revenge qua objection

The literature on truth and paradox exhibits a familiar and ubiquitous pattern: each proposed 'account of truth' is followed by a charge of *revenge*, that the account can't accommodate such and so a notion (e.g., 'untruth',

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

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

'exclusively false', or whathaveyou) and, in that respect, is thereby inadequate. Indeed, were it not for alleged 'revenge' problems, many proposed theories of truth might be objection-free—or, at least, the number of known or cited objections would be greatly diminished.

Such 'revenge' charges, as said, are often launched as inadequacy objections against proposed accounts of truth. Unfortunately, there is some unclarity about the relevance of such charges, and, more to the point, unclarity with respect to the burden involved in successfully establishing the intended inadequacy result. Without aiming to resolve them, §4 briefly discusses some of the given issues involved in *revenge qua objection*. Before turning to §4, two background issues need to be briefly covered.[7]

3.2.1 Incoherent Operators

By way of background, it is important to see that there are operators that cannot coherently exist if our language enjoys various features. Tarski's Theorem gives one concrete example of such a result,[8] but another example might be useful. In particular, suppose, as is plausible, that our language has features F1 and F2.

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

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

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

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

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

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

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

which yields two cases.

1. Case one:

   (a) $Tr(\ulcorner\lambda\urcorner)$

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

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

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

(b) Release yields:[10] $\Phi Tr(\ulcorner\lambda\urcorner)$.

(c) E2 yields: $\bot$

2. Case two:

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

(b) Capture yields: $Tr(\ulcorner\lambda\urcorner)$

(c) E2 yields: $\bot$

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

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

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

3.2.2 Models and Reality

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

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

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

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

1. We construct an artificial *model language*—one that's intended to serve as a heuristic, albeit idealized, model of our own 'real' language—and, in turn, give an account of how 'true' behaves in that language by constructing a precise account of *truth-in-that-language*.

2. We then claim that the behavior of 'true' in our language, at least in relevant, target respects, is like the behavior of the truth predicate in our model language.

By far the most dominant approach towards the first task—viz., constructing one's model language—employs a classical set theory. One reason for doing so is that classical set theory is familiar, well-understood, and generally taken to be consistent. A related reason is that, in using a classical set theory, one's formal account of truth can be more than merely a heuristic picture; it can also serve as a 'model' in the technical sense of *establishing consistency*.[12]

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

4 Comments on Revengers' Revenge

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

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

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

4.1 Too easy revenge

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

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

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

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

The revenger's charge, then, amounts to this: that the Kripkean model language fails to be enough like our real language to explain at least one of the target phenomena, namely, truth's consistency. Our metalanguage is

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

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

part of our 'real language', and we can define $\{\beta : \beta \text{ is not true-in-}\mathcal{L}_m\}$ in our metalanguage. As the Kripkean language cannot similarly define $\{\beta : \beta \text{ is not true-in-}\mathcal{L}_m\}$, the Kripkean model language is inadequate: it fails to illuminate truth's target features.

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

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

Such an argument might serve to turn otherwise 'too easy revenge' into a plainly relevant and powerful objection; however, the argument itself relies on various assumptions that involve quite complex issues. For example, one conspicuous assumption is that the 'semantics' of $\mathcal{L}_m$ is intended to reflect the semantics of our real language $\mathcal{L}$. This needn't be the case. For example, suppose that one rejects that semantics—the semantics of our real language—is a matter of giving 'truth conditions' or otherwise involves some explanatory notion of truth. In the face of Liars (or other paradoxes), one still faces questions about one's truth predicate, and in particular its logical behavior. By way of answering such questions, one might proceed as above: construct a model language that purports to illuminate how one's real truth predicate enjoys its relevant features (e.g., Capture and Release) without collapsing from paradox. In constructing and, in turn, describing one's 'model language', one might give 'truth-conditional-like semantics' for the model language by giving 'truth-in-a-model conditions' for the language. If so, it is plain that the 'semantics' of the model language are not intended to reflect the 'real semantics' of one's real language; they may, in the end, be only tools used for illuminating the logic of our real language, versus illuminating the 'real semantics' of our real language. So, a critical assumption in the argument above—the argument towards the relevance of the given inexpressibility results—requires argument. Likewise, the assumption that our real language can 'express its own semantic notions', the notions involved in 'giving the semantics' of our language, requires argument, argument that may turn, as with the first assumption, on difficult issues concerning the very 'nature of semantics'.[17]

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

[17] Some of the papers in this volume discuss this assumption, an assumption that often goes under the heading 'semantic self-

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

4.2 Revenger's Recipes, in general

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

Rv1. Recipe One.

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

Rv2. Recipe Two.

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

Rv3. Recipe Three.

- Find some semantic notion $\mathcal{X}$ that is (allegedly) in $\mathcal{L}$. (Argue that $\mathcal{X}$ is in $\mathcal{L}$.)
- Argue that $\mathcal{X}$ is not expressible in $\mathcal{L}_m$ lest $\mathcal{L}_m$ be inconsistent (or trivial).

sufficiency'. For arguments against such an assumption, see [7, 8].

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

– Conclude that $\mathcal{L}_m$ is explanatorily inadequate: it fails to explain how $\mathcal{L}$, with its semantic notion $\mathcal{X}$, enjoys consistency (or, more broadly, non-triviality).

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

The weight of Rv1 or Rv2 depends on the sort of $\mathcal{X}$ at issue. As in §3.2.2 and §4.1, if $\mathcal{X}$ is a classical, model-dependent notion constructed in a *proper fragment of* $\mathcal{L}$, then the charge of inadequacy is not easy to substantiate, even if the inexpressibility of $\mathcal{X}$ in $\mathcal{L}_m$ is easy to substantiate. In particular, if classical logic extends that of $\mathcal{L}_m$, then there is a clear sense in which you may 'properly' rely on a classical metalanguage in constructing $\mathcal{L}_m$ and, in particular, truth-in-$\mathcal{L}_m$. In familiar non-classical proposals, for example, you endorse that $\mathcal{L}$, the real, target language, is non-classical but enjoys classical logic as a (proper) extension, in which case, notwithstanding particular details, there is nothing prima facie suspect about relying on an entirely classical fragment of $\mathcal{L}$ to construct your model language and, in particular, classical model-dependent $\mathcal{X}$s. But, then, in such a context, it is hardly surprising that $\mathcal{X}$, being an entirely classical notion, would bring about inconsistency or, worse, triviality, in the (classically constructed) *non-classical* $\mathcal{L}_m$.[19]

Because classical logic is typically an extension of the logic of $\mathcal{L}_m$, the point above is often sufficient to blunt, if not undermine, a revenger's charge, at least if the given recipe is Rv1 and Rv2. As in §4.1, the revenger must establish more than the unsurprising result that a (usually classically constructed) model-dependent $\mathcal{X}$ is expressible in $M(\mathcal{L}_m)$ but not in $\mathcal{L}_m$; she must show the relevance of such a result, which might well involve showing that some non-model-dependent notion—some relevant 'absolute' notion—is expressible in $\mathcal{L}$ but, on pain of inconsistency (or non-triviality), inexpressible in $\mathcal{L}_m$. And this task brings us to Rv3.

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

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

For example, recall, from §3.2.1, the discussion of 'incoherent operators', and assume that Theorist proposes a theory that has features F1 and F2 (and, for simplicity, is otherwise normal with respect to extensional con-

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

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

The difficulty in successfully launching Rv3 might be put, in short, as follows. Theorist advances $\mathcal{L}_m$ as a model of (relevant features of) $\mathcal{L}$, our real language. Rv3 Revenger alleges that $\mathcal{X}$ exists in $\mathcal{L}$, and shows that, on pain of triviality, $\mathcal{X}$ is inexpressible in $\mathcal{L}_m$. The difficulty in adjudicating the matter is that, as in §3.2.1, Theorist may reasonably conclude that $\mathcal{X}$ is incoherent (given the features of our language that Theorist advances). Of course, if Revenger could establish that we *need* to recognize $\mathcal{X}$, perhaps for some theoretical work or otherwise, then the debate might be settled; however, such arguments are not easy to come by.

The burden, of course, lies not only on the Rv3 Revenger; it also lies with the given theorist. For example, typical paracomplete and paraconsistent theorists must reject the intelligibility of any EE device in our language. Inasmuch as such a notion is independently plausible—or, at least, independently intelligible—such theorists carry the burden of explaining why such a notion appears to be intelligible, despite its ultimate unintelligibility. Along these lines, the theorist might argue that we are making a common, reasonable, but ultimately fallacious generalization from 'normal cases' to all cases, or some such mistake. (E.g., some connective, if *restricted* to a proper fragment of our language, behaves in the EE way.) Alternatively, such theorists might argue that, contrary to initial appearances, the allegedly intelligible notion only appears to be a clear notion but, in fact, is rather unclear; once clarified, the alleged EE device (or whatever) is clearly not such a device. (E.g., one might argue that the alleged notion is a conflation of various notions, each one of which is intelligible but not one of which behaves in the alleged, problematic way.) Whatever the response, theorists do owe something to Rv3 revengers: an explanation as to why the given (and otherwise problematic) notion is unintelligible.

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

5 Some closing remarks

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

Chapter Summaries

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

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

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

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

Rayo–Welch).[21]

Hofweber. Validity is often thought to be truth-preserving: an inference rule is valid just if truth-preserving.[22] Thomas Hofweber, in his 'Validity, Paradox, and the Ideal of Deductive Logic', argues that two senses of 'an inference rule is valid just if truth-preserving' are important to distinguish. One sense is the 'strict reading', according to which *each and every* instance of the given rule is truth-preserving. The other reading is the 'generic reading', which, in some sense, is analogous to the claim that *bears are dangerous*, a claim that is true even though not true of all bears. This distinction, which Hofweber discusses, holds the key to resolving the revenge phenomenon. In particular, the Liar's revenge teaches us that we should abandon the traditional ideal of deductive logic, which requires that our theories be underwritten by rules that are valid in the 'strict sense'. On the positive side, the Liar's revenge teaches us that we should embrace the 'generic' ideal of deductive logic, which requires only that our rules be 'generically valid'.

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

Maudlin. In his 'Reducing Revenge to Discomfort', Tim Maudlin argues that the revenge phenomenon ultimately teaches us something about our normative principles of assertion. As in §3 (above), invoking a new category for Liars—say, *bugger*—seems inevitably to lead to new Liars (e.g., the starred sentence in §3 above). While Maudlin maintains that we do need three semantic categories (viz., truth, falsity, and ungroundedness), he argues that

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

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

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

we need no more than three. In particular, we may—and should—assert that the ticked sentence in §2 above is not true; it's just that we'll be bucking the traditional principle according to which only truths are properly assertible. The problem, Maudlin argues, is not with principles of truth (e.g., Release and Capture); the problem is with the traditional principle of assertion.[24] On the other hand, Maudlin admits that the revenge phenomenon returns even for his revised principle of assertion (e.g., 'I am not properly assertible according to Maudlin's revised principles'). Maudlin argues that this is revenge, but that it is at most a discomfort; it is far from threatening the coherence of *truth*.

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

Priest. In his 'Revenge, Field, and ZF', Graham Priest does three things. First, Priest characterizes the Liar's revenge, and carves up three options for dealing with it. Second, Priest directs the discussion towards Field's chapter (see Chapter ??), and argues that Field's proposal is not revenge-free, contrary to Field; in particular, it faces an expected problem with the notion of *having value* 1. (Priest anticipates the immediate thought that, as sketched in §4.1 above, he is merely launching a form of 'too-easy revenge', conflating model-dependent and 'real' notions. Priest argues that unless 'having value 1' is a real notion, Field has given no reason to think that Field's proposed logic has anything to do with real validity—i.e., validity in our real language.) Third, Priest argues that the (alleged) troubles facing Field's proposal are a symptom of deeper revenge in the background theory of ZF, which theory, Priest argues, itself faces a serious revenge-like situation involving V (the cumulative hierarchy): the logic defined by the theory (in terms of models) does not apply to the theory itself, thereby leaving us 'bereft of a justification for reasoning about sets', as Priest puts it.

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

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

theory (a language enjoying transparent truth and a suitable conditional). Such a notion, as Rayo & Welch argue, plays the Liar's revenge role: it would generate inconsistency were it expressible in Field's proposed language.[25]

Read. In his 'Bradwardine's Revenge', Stephen Read discusses a theory of truth proposed by Thomas Bradwardine (who was principally a physicist and theologian in the 1300s). Read shows that Bradwardine's theory, according to which Liars are not true (because they'd have to be true and not true, which is impossible), is a subtler theory than the later Buridan-like theories that, in effect, reject unrestricted Capture for truth (see §1 above). Moreover, the theory, on the surface, as Read argues, seems to promise a revenge-free approach to a whole host of semantic paradoxes. The key for Bradwardine is to distinguish between the claim that the Liar is false from the Liar itself. The propositions appear to be indistinguishable, but they are not. According to Bradwardine, any proposition that 'says' of itself that it is false, also 'says' of itself that it is true. (As Read points out, this is a subtler thesis than the later Buridanian claim that every claim 'says' of itself that it is true.)

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

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

Shapiro. In his 'Burali-Forti's Revenge', Stewart Shapiro turns the focus from the Liar paradox to the Burali-Forti paradox, which, he argues, has its own revenge issues. (Using the later von Neumann account, which came

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

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

after Burali-Forti, the paradox, in short, is that the set Ω of all ordinals satisfies all that's required to be an ordinal, in which case, the successor of Ω, namely $\Omega + 1$, is strictly greater than Ω. But, being itself an ordinal, $\Omega + 1$ must be in Ω, giving the result that $\Omega < \Omega + 1 \leq \Omega$, which is impossible.) Shapiro presents the paradox and a variety of ways of dealing with it. He argues that each option faces severe problems, leaving the matter open.

Simmons. In his 'Revenge and Context', Keith Simmons first distinguishes between (what he calls) *direct revenge* and *second-order revenge*. The former variety is the (what one might call 'first-order') variety: we already have a stock of semantic terms, and they generate paradox. In particular, as with the ticked sentence in §3, one is naturally inclined to classify it as 'neither true nor false', but this (at least prima facie) implies untruth, and the paradox remains. One is stuck in direct revenge: an inability to classify the sentence as one thinks it ought to be classified—but cannot be so classified, on pain of inconsistency. But, now, one introduces new, technical machinery to deal with the direct revenge problem: one calls the Liar a *bugger*, or *unstable*, or *whatnot*. Second-order revenge emerges with this new machinery, and one is, again, unable to classify the (new) Liars as they 'ought' to be (in some sense). Simmons argues that, while there is still work to be done, his 'singularity theory' of semantic notions deals not only with direct revenge in a natural way; it also holds the promise of resolving second-order revenge.

* * *

Appendix

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

Philosophical picture

One conception of truth has it that truth is entirely *transparent*, that is, a truth predicate $Tr(x)$ in (and for) our language such that $Tr(\ulcorner\alpha\urcorner)$ and α are intersubstitutable in all (non-opaque) contexts, for all α in the language. This conception comes with a guiding metaphor, according to which 'true' is introduced only for purposes of generalization. Prior to introducing the device, we spoke only the 'true'-free fragment. (Similarly for other semantic notions/devices, e.g., 'denotes', 'satisfies', 'true of', etc.) For simplicity, let us assume that the given

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

'semantic-free' fragment (hence, 'true'-free fragment) is such that LEM holds.[28] Letting $\mathcal{L}_0$ be our 'semantic-free fragment', we suppose that $\alpha \vee \neg\alpha$ is true for all α in $\mathcal{L}_0$.[29] Indeed, we may suppose that classical semantics—and logic, generally—is entirely appropriate for the fragment $\mathcal{L}_0$.

But now we want our generalization-device. How do we want this to work? As above, we want $Tr(\ulcorner\alpha\urcorner)$ and α to be intersubstitutable for *all* α. The trouble, of course, is that once 'is true' is introduced into the language, various unintended—and, given the role of the device, paradoxical—sentences emerge (e.g., the ticked sentence in §2 above).[30]

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

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

1. 'Max is a cat' is true

the guiding principle instructs that you may also assert

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

And so on. More generally, your learning can be seen as a process of achieving further and further truth-attributions to sentences that 'ground out' in $\mathcal{L}_0$. (Similarly for falsity, which is just truth of negation.) Eventually, your competence reflects precisely the defining intersubstitutivity—and transparency—of truth: that $Tr(\ulcorner\alpha\urcorner)$ and α are intersubstitutable for *all* α of the language.

But your competence also reflects something else: namely, the failure to assert either α or $\neg\alpha$, for some α in the language. To see the point, think of the above process of 'further and further truth-attributions' as a process of writing two (very, very big) books—one, *The Truth*, the other *The False*. Think of each stage in the process as

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

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

[30] With respect to formal languages, the inevitability of such sentences is enshrined in Gödel's so-called *diagonal lemma*. (Even though the result is itself quite significant, it is standardly called a *lemma* because of its role in establishing Gödel–Tarski indefinability theorems. For user-friendly discussion of the limitative results, and for primary sources, see [18]. For a general discussion of diagonalization, see [9].)

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

completing a 'chapter', with chapter zero of each book being empty—this indicating that *at the beginning* nothing is explicitly recorded as true (or, derivately, false).

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

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

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

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

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

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

So goes the basic process for *atomics*. But what about compound (molecular) sentences? The details are sketched below (see 'Formal model'), but for now the basic idea is as follows (here skipping the relativizing to chapters). With respect to negations, $\neg\alpha$ goes into *The True* just when α goes into *The False*. (Otherwise, neither α nor $\neg\alpha$ finds a place in either book.) With respect to *conjunctions*, $\alpha \wedge \beta$ goes into *The False* if either α or β goes into *The False*, and it goes into *The True* just if both α and β go into *The True*. (Otherwise, $\alpha \wedge \beta$

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

finds a place in neither book.) The case of *disjunctions* is dual, and the quantifiers may be treated as 'generalized conjunction' (universal) and 'generalized disjunction' (existential). This approach to compound sentences reflects the so-called *Strong Kleene* scheme, which is given below (see 'Formal model').

Does every sentence eventually find a place in one book or other? No. Consider an atomic sentence λ, like the ticked sentence in §2, equivalent to $\neg Tr(\ulcorner\lambda\urcorner)$. In order to get λ into *The True* book, there'd have to be some chapter in which it appears. λ doesn't appear in chapter zero, since nothing does. Moreover, λ doesn't exemplify anything 'according to chapter zero', since chapter zero concerns only the $\mathcal{L}_0$-sentences (and λ isn't one of those). What about chapter one? In order for λ to appear in chapter one, λ would have to be in chapter zero or be such that λ exemplifies $\neg Tr(x)$ according to chapter zero. But for reasons just given, λ satisfies neither disjunct, and so doesn't appear in chapter one. The same is evident for chapter two, chapter three, and so on. Moreover, the same reasoning indicates that L doesn't appear in *The False* book.

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

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

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

Formal model

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

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

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

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

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

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

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

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

that, for any predicate, the extension and antiextension are jointly exhaustive (the union of the two equals the domain) and, of course, exclusive (the intersection of the two is empty).

Concentrating on the so-called *Strong Kleene* account [10],[36] the formal story runs as follows. We expand $\mathcal{V}$, as above, to be $\{1, \frac{1}{2}, 0\}$, and so our language $\mathcal{L}_\kappa = \langle \mathtt{L}, \mathcal{M}, \kappa \rangle$ is now a so-called three-valued language (because it uses three semantic values).[37] Our *designated values*—intuitively, the values in terms of which *validity* or *consequence* is defined—are a subset of our semantic values; in the Strong Kleene case, there is exactly one designated element, namely 1.

A (Strong Kleene) model $\mathcal{M} = \langle \mathcal{D}, I \rangle$ is much as before, with I doing exactly what it did in the classical case except that I now assigns to n-ary predicates elements of $D^n \longrightarrow \{1, \frac{1}{2}, 0\}$, since $\mathcal{V} = \{1, \frac{1}{2}, 0\}$. Accordingly, the 'intensions' of our paracomplete (Strong Kleene) predicates have three options: 1, $\frac{1}{2}$, and 0. What about *extensions*? As above, we want to treat predicates not just in terms of extensions (as in the classical languages) but also antiextensions. The *extension* of an n-ary predicate F, just as before, comprises all n-tuples $\langle a_1, \ldots, a_n \rangle$ of $\mathcal{D}$ such that $I(F)(\langle a_1, \ldots, a_n \rangle) = 1$. (Again, intuitively, this remains the set of objects of which F is true.) The *antiextension*, in turn, comprises all n-tuples $\langle a_1, \ldots, a_n \rangle$ of $\mathcal{D}$ such that $I(F)(\langle a_1, \ldots, a_n \rangle) = 0$. (Again, intuitively, this is the set of objects of which F is false.) Of course, as intended, an interpretation might fail to put x in either the extension or antiextension of F. In that case, we say (in our 'metalanguage') that, relative to the model, F is *undefined* for x.[38]

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

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

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

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

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

K1. $V_{\mathcal{M}}(\neg\alpha) = 1 - V_{\mathcal{M}}(\alpha)$.

K2. $V_{\mathcal{M}}(\alpha \vee \beta) = \mathtt{max}(V_{\mathcal{M}}(\alpha), V_{\mathcal{M}}(\beta))$.

K3. $V_{\mathcal{M}}(\exists x\, \alpha(x)) = \mathtt{max}\{V_{\mathcal{M}}(\alpha(t/x)) :$ for all $t \in \mathcal{D}\}$.

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

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

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

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

The construction runs (in effect) along the lines of the 'big books' picture. For simplicity, let $\mathcal{L}_{\kappa}$ be a classical (but nonetheless Strong Kleene) language such that `L` (the basic syntax, etc.) is free of semantic terms but has the resources to describe its given syntax—including, among other things, having a name $\ulcorner\alpha\urcorner$ for each sentence α. (In other words, I assigns to each n-ary predicate an element of $\mathcal{D}^n \longrightarrow \{1, 0\}$, even though the values $\mathcal{V}$ of

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

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

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

$\mathcal{L}_\kappa$ also contain $\frac{1}{2}$.) What we want to do is move to a richer language the syntax $\mathtt{L}^t$ of which contains $Tr(x)$, a unary predicate intended to be a transparent truth predicate for the enriched language. For simplicity, assume that the domain $\mathcal{D}$ of $\mathcal{L}_\kappa$ contains all sentences of $\mathtt{L}^t$.[42]

Think, briefly, about the 'big books' picture. One can think of each successive 'chapter' as a language that expands one's official record of what is true (false). More formally, one can think of each such 'chapter' of both books as the extension and antiextension of 'true', with each such chapter expanding the interpretation of 'true'. Intuitively (with slight qualifications about chapters zero), one can think of $I_{i+1}(Tr)$ as explicitly recording *what is true according to chapter* $I_i(Tr)$. The goal, of course, is to find a 'chapter' at which we have $I_{i+1}(Tr) = I_i(Tr)$, a 'fixed point' at which anything true in the language is fully recorded in the given chapter—one needn't go further. Thinking of the various 'chapters' as *languages*, each with a richer interpretation of 'true', one can think of the 'fixed chapter' as a language that, finally, has a transparent truth predicate for itself.

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

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

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

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

of 'is true'. As above, letting $I_i(Tr)$ be a function $(\mathcal{T}_i^+, \mathcal{T}_i^-)$ yielding 'both chapters i', the goal is to eventually 'jump' upon an interpretation $(\mathcal{T}_i^+, \mathcal{T}_i^-)$ such that $(\mathcal{T}_i^+, \mathcal{T}_i^-) = (\mathcal{T}_{i+1}^+, \mathcal{T}_{i+1}^-)$.

Focusing on the 'least such point' in the Strong Kleene setting, Kripke's construction proceeds as above. We begin at stage 0 at which $Tr(x)$ is interpreted as $(\emptyset, \emptyset)$, and we define a 'jump operator' on such interpretations:[44] $Tr(x)$ is interpreted as $(\mathcal{T}_{i+1}^+, \mathcal{T}_{i+1}^-)$ at stage $i+1$ if interpreted as $(\mathcal{T}_i^+, \mathcal{T}_i^-)$ at the preceding stage i, where, note well, $\mathcal{T}_{i+1}^+$ comprises the sentences that are true (designated) at the preceding stage (chapter, language) i, and $\mathcal{T}_{i+1}^-$ the false sentences (and, for simplicity, non-sentences) at i. Accordingly, we define the 'jump operator' J_{SK} thus:[45]

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

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

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

defined (via transfinite recursion) thus:[46]

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

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

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

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

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

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

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

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

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

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

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

References

[1] JC Beall. 'True, false, and paranormal'. *Analysis*, 66(2):102–114, 2006. Available in *Analysis Preprint* series.

[2] JC Beall. 'Truth and paradox: a philosophical sketch'. In Dale Jacquette, editor, *Philosophy of Logic*, *Handbook of Philosophy of Science*, under the general editorship of Dov Gabbay, Paul Thagard, and John Woods.), pages 187–272. Elsevier, Dordrecht, 2006.

[3] Solomon Feferman. 'Toward Useful Type-Free Theories, I'. *Journal of Symbolic Logic*, 49:75–111, 1984. Reprinted in [12].

[4] Hartry Field. 'The Semantic Paradoxes and the Paradoxes of Vagueness'. In JC Beall, editor, *Liars and Heaps: New Essays on Paradox*, pages 262–311. Oxford University Press, Oxford, 2003.

[5] Hartry Field. 'Truth and the Unprovability of Consistency'. *Mind*, 115:567–606, 2006.

[6] Melvin Fitting. 'Notes on the mathematical aspects of Kripke's theory of truth'. *Notre Dame Journal of Formal Logic*, 27:75–88, 1986.

[7] Anil Gupta. 'Definition and Revision'. In Enrique Villanueva, editor, *Truth*, number 8 in Philosophical Issues, pages 419–443. Ridgeview Publishing Company, Atascadero, California, 1997.

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

[9] Dale Jacquette. 'Diagonalization in Logic and Mathematics'. In Dov M. Gabbay and Franz Günthner, editors, *Handbook of Philosophical Logic*, pages 55–147. Kluwer Academic Publishers, Dordrecht, Second edition, 2004.

[10] S. C. Kleene. *Introduction to Metamathematics*. North-Holland, 1952.

[11] Saul Kripke. 'Outline of a Theory of Truth'. *Journal of Philosophy*, 72:690–716, 1975. Reprinted in [12].

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

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

[14] Graham Priest. *Doubt Truth To Be A Liar*. Oxford University Press, Oxford, 2006.

[15] Graham Priest. *In Contradiction*. Oxford University Press, Oxford, Second edition, 2006.

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

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

[16] William N. Reinhardt. 'Some remarks on extending and interpreting theories with a partial predicate for truth'. *Journal of Philosophical Logic*, 15:219–251, 1986.

[17] Brian Skyrms. 'Return of the Liar: Three-Valued Logic and the Concept of Truth'. *American Philosophical Quarterly*, 7:153–161, 1970.

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

[19] Raymond M. Smullyan. *Recursion Theory for Metamathematics*, volume 20 of *Oxford Logic Guides*. Oxford University Press, New York, 1993.

[20] Raymond M. Smullyan. 'Gödel's Incompleteness Theorems'. In Lou Goble, editor, *The Blackwell Guide to Philosophical Logic*, pages 72–89. Blackwell, Oxford, 2001.

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

[22] Albert Visser. 'Semantics and the liar paradox'. In Dov M. Gabbay and Franz Günthner, editors, *Handbook of Philosophical Logic*, pages 149–240. Kluwer Academic Publishers, Dordrecht, Second edition, 2004.

[23] Stephen Yablo. 'New Grounds for Naïve Truth Theory'. In JC Beall, editor, *Liars and Heaps: New Essays on Paradox*, pages 313–330. Oxford University Press, Oxford, 2003.