To appear in B. Armour-Garb (ed.), *The Relevance of the Liar* (Oxford University Press)

Truth, Hierarchy and Incoherence

Bruno Whittle

Approaches to truth and the Liar paradox seem invariably to face a dilemma. Either they must eventually appeal to some sort of hierarchy, or else they must say that some apparently perfectly coherent concept is in fact *in*coherent. But each option would seem to involve severe expressive restrictions, and would thus seem unsatisfactory. The aim of this paper is a new approach, which avoids the dilemma and the expressive restrictions. Previous approaches often try to solve the Liar by appealing to some sort of new semantic value for the truth predicate to take; but I will argue that approaches of this sort inevitably face the dilemma in question. In contrast, the approach that I will propose will stick to classical semantic values, but will allow that the compositional rules associated with these have exceptions. One might worry that it will be impossible to develop such an approach rigorously and systematically—dispensing, as it does, with exceptionless compositional rules. But I will explain how such a development is in fact possible.

The structure of the paper is as follows. In §1 I will explain how existing approaches tend to face the dilemma in question, and why neither horn would seem satisfactory. In §2 I will outline the proposed approach and explain why—*if* it could be adequately developed—it would seem to promise a way out of the dilemma. In §3 I will explain how the approach can be adequately developed. In §4 I will consider objections. And in §5 I will discuss the implications for logic, if the approach is correct.

1. The Hierarchy-Incoherence Dilemma

The aim of this section is thus to describe a general dilemma that approaches to truth tend to suffer from, and to explain why neither horn would seem to be satisfactory.[1] For the purposes of illustration, I will focus on the approach of Kripke [1975], but I will also indicate how a range of other approaches face the dilemma.

[1] For discussions of issues closely related to the dilemma that I will raise see, for example, Kripke [1975: 714–15], Maudlin [2004: 26–67] and Field [2008: 309–24].

The aim of the approach of Kripke [1975] is to give an account of how languages can contain their own truth predicates—despite the existence of (for example) sentences that say of themselves that they are not true. The basic idea is to give the truth predicate a new sort of semantic value. Thus, rather than a simple extension, the truth predicate is assigned a pair of disjoint sets: an extension together with an 'anti-extension', where the idea is that the predicate is true of the things in its extension, false of the things in its anti-extension, and neither true nor false of everything else. This new sort of semantic value for the truth predicate necessitates an extended approach to the logical operators (i.e., the connectives and the quantifiers). For suppose that T is the truth predicate, and that Tc is neither true nor false. What then about ¬Tc, for example? Is this sentence true, false or neither? In fact, Kripke's general approach works for a number of different 'evaluation schemes' (i.e., a number of different extended approaches to the logical operators). But the approach that he focuses on—and which commentators have followed him in focussing on—is the 'strong Kleene' scheme; so I will similarly focus on this scheme here.[2] Thus, use u (for 'undefined') for the value taken by sentences that are neither true nor false. Then, according to the strong Kleene scheme, ¬ is interpreted so as to correspond to the function that sends t to f, f to t, and u to itself: so ¬A is true if A is false, false if A is true, and neither if A is neither.[3]

This understanding of ¬ allows languages to contain their own truth predicates—despite the presence of Liar sentences. For, on this approach, if T is a truth predicate for a language, then its extension will be the set of true sentences of the language. Thus, suppose that T is a truth predicate for a language that itself contains both T and a sentence ¬Tc, where c denotes this very sentence (which I will call B). B clearly cannot be either true or false: if it was true, then (by the strong Kleene understanding of ¬) Tc would be false; i.e., the denotation of c (= B) would be in the anti-extension (and not the extension) of T; but B being true then contradicts the hypothesis that T is a truth predicate for the language; similarly, if B is false then (by the strong Kleene treatment of ¬ again) it

[2] However, everything that I say could easily be modified so as to apply to any of the alternative versions of the approach. For these see Kripke [1975: 711–12].

[3] The scheme handles other operators as follows. A conjunction is true if both of the conjuncts are, false if one of them is false, and neither otherwise. A sentence of the form ∀xA(x) is true if A(x) is true of everything in the domain, false if it is false of something in the domain, and neither otherwise. Other operators are defined out of negation, conjunction and universal quantification in the usual way.

follows that B *is* in the extension of T, and this again contradicts the hypothesis that T is a truth predicate for the language (since, by hypothesis, B is false, and so *not* true). But, on the other hand, we do not get any sort of contradiction if we instead say that B is neither true nor false: in this case all we get from the strong Kleene understanding of ¬ is that Tc is *also* neither true nor false, and thus that the denotation of c (= B) is in neither the extension nor the anti-extension of T; but this of course in no way contradicts T being a truth predicate for the language. Thus, this understanding of negation seems to allow for languages that contain their own truth predicates, even if they also contain Liar sentences.[4]

To see the limitations of this approach, however, one need only consider a simple alternative to the strong Kleene understanding of negation. For, once one allows predicates to have pairs of sets as their semantic values, there is then more than one natural treatment of negation. Thus, strong Kleene negation corresponds to the function that sends t to f, f to t and u to u. But a very natural alternative is 'exclusion' negation, which (corresponds to the function that) sends t to f, f to t, and u to t. So use ~ for this new sort of negation (and continue to use ¬ for strong Kleene negation). To see why this is a natural form of negation, suppose for example that T is a truth predicate, and that its semantic value is $\langle X,Y \rangle$. Then something is true (in the sense of T) iff it is X. Thus, something is not true (in this sense) iff it is *not* in X. But to express this notion of 'not true' (or of 'untruth') one needs exclusion negation: for ~Tx is true of everything (in the domain) that is not in X, whereas ¬Tx is true only of those things that are *both* not in X and *also* in Y.

The problem, however, is that the approach completely breaks down once exclusion negation is added to the language. For suppose that T is a truth predicate for some language that contains exclusion negation, and let ~Tc be a sentence of this language, such that c denotes this very sentence (call this sentence E).[5] E is then either

[4] For further details of the approach see Kripke [1975].

[5] Throughout I will assume that the languages discussed contain self-reference, in the sense that for any formula of the language A(x), there will be a closed term of the language b denoting the sentence A(b). The reason for doing this is that any viable approach to truth must work even in the presence of self-reference: for even if one was to prohibit 'direct' self-reference (e.g., using 'Jack' to name the sentence 'Jack is short'), still there could be no hope of prohibiting 'indirect' self-reference (e.g., sentences such as 'Every sentence uttered by the lottery winner is short' uttered by the lottery winner). It is then merely a harmless

true or false (every sentence of the form ~A is, because ~ sends each of t, f and u to t or f). But if it is true, then the denotation of c (= E) must *not* be in the extension of T: contradicting the hypothesis that T is a truth predicate for the language. But, similarly, if E is false, then it must be *in* the extension of T: meaning, again, that T is not, after all, a truth predicate for the language.

Thus, exclusion negation presents the approach with a dilemma. One option is to accept that if one starts with a language that contains exclusion negation, then, to talk about truth in that language one must use an *extension* of this language; but then, to talk about truth in the extended language, one must use yet another extension, and so on—leading to a hierarchy of languages. And the only alternative to this would seem to be to claim that exclusion negation is in some sense an illegitimate, or 'incoherent', notion (i.e., reject the notion, and with it the need for a hierarchy).

The problem, however, is that neither option would seem to be satisfactory. Consider first the possibility of appealing to a hierarchy of languages, once exclusion negation is in the picture. I will mention just one drawback of this option. This is that if languages that contain exclusion negation cannot contain their own truth predicates, then it would seem that no language can contain a *general* truth predicate (i.e., a predicate that applies to true sentences generally, rather than merely those that belong to some particular language). For suppose for the sake of argument that L is such language. Then either L already contains exclusion negation, or it can surely be extended to such a language. But a language that contains both exclusion negation and a general truth predicate is, in effect, a language that contains both exclusion negation and its *own* truth predicate—which is impossible on this approach.[6] Thus, on this horn of the dilemma, it would seem that there can be no general truth predicates. But this would be a severe expressive restriction. For without such a predicate it will be impossible to express general claims about language such as 'Nothing is both true and false', or general norms governing communication such as 'One should only assert what is true'. But it is surely

---

simplification to assume that all of the languages discussed already contain self-reference, and that in the sense described.

[6] More precisely, the argument that showed that (on this approach) languages cannot contain both exclusion negation and their own truth predicates can easily be modified to show that (on this approach) no language can contain both exclusion negation and a general truth predicate.

essential to philosophy that we be able to express such things, and so this horn of the dilemma would seem to be unsatisfactory.

What about the second option, of saying that exclusion negation is incoherent? The problem is that this option *also* seems to involve severe expressive restrictions. For even if this option allows for general truth predicates, it does so only at the cost of a restriction on what one can say using them. For although one could (on this option) perhaps express the general claim 'Nothing is both true and false', there will be other very natural claims that one will *not* be able to express. For example, the version of the claim 'Everything is either true or not true' that uses exclusion negation (and I hope that I have already made clear just how natural a version of this claim this is). Thus this option *also* seems to lead to severe expressive restrictions, and would thus also seem to be unsatisfactory. Hence, overall, the problem that the dilemma raises for Kripke's approach.[7]

And a similar dilemma (and hence a similar problem) would seem to be faced by a wide range of other approaches to truth. For reasons of space, I will not discuss any of these in detail. But, in the remainder of this section, I will indicate how a number of these face a version of the dilemma.

Thus, consider first the 'revision theory' of truth (see, e.g., Gupta [1982], Herzberger [1982] and Gupta and Belnap [1993]). On this approach, truth predicates have 'revision rules' as their semantic values: i.e., functions from 'hypothetical' extensions to 'revised' extensions. As in the case of Kripke's approach, however, there will—on this approach—be natural forms of negation that languages cannot contain in addition to their own truth predicates. For example, a form of negation ~ such that ~Tc is true (under a hypothetical extension for T) iff the denotation of c does not belong to *any* revised extension for T.[8] The upshot is that the revision theory faces a dilemma similar to that faced by Kripke's approach, and where each option would seem to be similarly unsatisfactory.

---

[7] I should note that Kripke himself seems clearly to prefer the former horn (i.e., 'hierarchy' rather than 'incoherence'): see [1975: 714–15].

[8] To spell out the argument, let r be the revision rule for T. Then, for a hypothetical extension X, r(X) is the set of sentences that are true, under the hypothesis that X is the extension of T. But then consider a sentence ~Tb, where b denotes this very sentence (call it A). Then for any X, $A \in r(X)$ iff for any Y, $A \notin r(Y)$ (by the definition of ~). So $A \notin r(X)$. But then $A \in r(X)$ after all: and so we have a contradiction. Thus, as claimed, the approach cannot handle languages that contains both ~ and their own truth predicates.

Two more recent approaches are those of Maudlin [2004] and Field [2008]. Maudlin's approach is closely related to Kripke's,[9] and so exclusion negation results in a similar dilemma for this approach. However, a significant part of Maudlin [2004] is devoted to *defending* the incoherence-horn; specifically, to arguing that there are independent reasons for thinking that exclusion negation is incoherent (see pp. 26–67). Put simply, the basic idea is that no legitimate connective can send u to either t or f, because if a sentence A receives u, then the 'facts' do not determine whether A is the case; but this 'determinacy failure' cannot then *determine* that another sentence (e.g., ~A) *is* true (say)—because, essentially, truth or falsity can only be determined by 'facts', and not by the mere absence of these. Intuitively, however, it is hard to be completely convinced by this: for language would seem to be a pretty flexible tool, and it is hard to see anything truly incoherent in the idea of one sentence being made true simply by another *failing* to be so made. I would suggest, then, that even given Maudlin's defence, his approach would seem to lead to an unsatisfactory pair of options.

The approach of Field [2008] does not propose a new semantic value for the truth predicate to take.[10] But it does something sufficiently similar that it seems to face a version of the dilemma similar to those discussed above. Thus, a cornerstone of this approach is the rejection of the law of excluded middle for sentences of the form Tc: i.e., the rejection of (some) sentences of the form Tc ∨ ¬Tc. Intuitively, it would thus seem that a sentence can fail to be true (in the sense of T) but still not be 'untrue' in the sense of ¬T (as in cases in which the relevant instance of excluded middle is rejected). But surely we can *also* introduce a form of negation ~ such that a sentence ~Tc will be true precisely when the denotation of c simply fails to be true (in the sense of T). However, Field's approach cannot handle languages that contain *both* such an operator *and* their

[9] Although there are significant differences between the two approaches. For example, Maudlin [2004] contains a distinctive account of the permissibility of assertions (where permissibility radically diverges from truth; see especially pp. 95–104 and 141–77).

[10] A component of this approach *is* a model-theoretic semantics for languages that contain their own truth predicates. However, this is claimed merely to establish the consistency of the theory, and not to provide a 'model' (in the informal sense) of a new sort of semantic value.

own truth predicates, and so one would seem to once again to face a version of the dilemma.[11]

Thus, a range of approaches would seem to face the 'hierarchy-incoherence dilemma': either they must eventually appeal to some sort of hierarchy of languages, or else they must say that some simple, and apparently perfectly coherent, concept is in fact *in*coherent. But, as we have seen, each option would seem to involve severe expressive restrictions, and so to be unsatisfactory. The aim of the rest of the paper is thus an approach that avoids the dilemma.

2. Avoiding the Dilemma: General Strategy

The approaches that we considered in the last section tried to solve the Liar by appealing to some new sort of semantic value for the truth predicate to take (or something similar to this); but, in each case, this led to a version of our dilemma. Further, it is plausible that any approach of this general sort will have just this problem: for it would seem that, for any new sort of semantic value (or similar), there will be a variety of negation that causes a problem for it, akin to the problem that exclusion negation causes for pairs of extensions and anti-extensions (in the case of Kripke's approach). What I want to suggest, then, is that if we want to avoid this problem, then we are going to need to try something fundamentally different. In particular, what I suggest is this: we should try to

---

[11] The argument can be spelt out as follows. If ~ is introduced as in the text, then it would seem that it must satisfy the following rules of inference (here $\perp$ can be taken either to be a logical constant, or a given contradiction, such as 0 = 1; and |= stands for logical consequence):

(1) A, ~A |= $\perp$, and
(2) if A |= $\perp$ then |= ~A.

For, suppose for the purposes of illustration that A is of the form Tc. Then it is surely impossible for the denotation of c to *both* be true in the sense of T, and also to fail to be; but then (1) must hold, given how we have introduced ~. Similarly, for (2), suppose that Tc |= $\perp$ holds; then the denotation of c must surely *fail* to be true in the sense of T; i.e., ~Tc must hold, as required. Thus, given how we have introduced ~, it would seem to be self-evident that rules (1) and (2) must hold for it.

However, if so, then it follows that no language can contain both its own truth predicate and ~, on Field's approach. For consider a sentence ~Ta where a denotes this very sentence. Then Ta |= $\perp$ (because, on Field's approach, one has Tb =||= B, where B is the denotation of b; and so Ta |= ~Ta; but then we have Ta |= $\perp$ by (1)). But then |= ~Ta (by (2)). And so we have |= Ta and thus |= $\perp$ (by (1)): contradiction. It would seem therefore that Field's approach faces a version of the dilemma similar to those discussed above.

stick to classical semantic values, but accept that the compositional rules associated with these have exceptions (for example, in cases such as the Liar). Thus, the idea would be that in the case of a Liar sentence ¬Tc (for example) we escape contradiction not by giving a new sort of semantic value to T, but—*rather*—by accepting that this sentence is an exception to the standard compositional rules; that is, we accept that this sentence is not true, despite the fact that if the compositional rules *did* apply to it, then (given the semantic values of ¬, T and c) they would give the result that it is *true*.

By the 'standard compositional rules' I mean the rules that determine the truth-value of a sentence on the basis of the semantic values of its atomic expressions that are enshrined in classical model theory. That is, the rules that say that a sentence ¬Ga is true (false) if the semantic value of a does not (does) belong to the semantic value of G; and a sentence ∃x(Gx ∧ Hx) is true (false) if some (no) object in the domain is in the semantic value of G and the semantic value of H; and so on. The proposal, then, is that in cases such as the Liar these rules have exceptions.

However, a very natural worry to have at this point is that it will be impossible to develop any such approach rigorously and systematically. For (one might think) surely any such development will require *exceptionless* rules—whereas the whole idea behind the proposed approach is to dispense with these. I will argue below, however, that we can in fact give a rigorous and systematic such development. Before doing that, however, I will explain why, if such a development *were* possible—i.e., if this sort of an approach *could* be adequately developed—then that would seem to promise a way out of the dilemma.

In fact, though, it is actually pretty clear why this is so. To see the point, recall Kripke's approach, for example. We forced this approach into the dilemma by considering exclusion negation ~ and a Liar sentence ~Tc. In particular, the hypothesis that T is a truth predicate for a language that contains this sentence led to contradiction. *But*: the derivation of this contradiction made essential use of the compositional rules associated with the semantic values of ~, T and c; for example, of the fact that if the denotation of c is not in the extension of T, then ~Tc will be true. Thus if we could develop an approach on which the compositional rules *fail* in such cases, then this general

sort of argument would be blocked. And so the sort of approach that I have described would seem to offer a way out of the dilemma—if it could be adequately developed.

However, as I noted, there does seem to be a prima facie concern about the possibility of such a development. Thus, I now move on to making the case that this really is possible. In fact, I will argue that we are already in possession of formal theories of truth that—although they were designed with very different aims in mind—can be seen to yield natural models of the sort of languages that I have described (i.e., languages with classical semantic values, but where the compositional rules associated with these have exceptions).[12] In particular, the formal theories that I have in mind were developed to respect what is sometimes called the 'Chrysippus intuition', which is as follows. Suppose that at time t Zeno says, 'What Zeno says at t is not true'. Then, according to this 'intuition', Zeno's utterance is neither true nor false. But—the intuition continues—if Chrysippus, having recognized this fact, says, 'What Zeno says at time t is not true', then *Chrysippus's* utterance is true—despite the fact that he has used exactly the same words that Zeno did, arranged in exactly the same way. That is the content of the 'intuition'.

I will discuss specific formal theories based on this intuition in a moment—in particular, using one such theory as an illustration, I will explain how these theories yield natural models of languages whose semantic values are classical, but where the compositional rules associated with these have exceptions. But I will first explain in general terms why such formal theories would seem to be the natural place to look for models of such languages. Thus, on any such theory, it will be possible for there to be two tokens of the form ¬Tc, where c names the first of these, and where the first token is neither true nor false, while the second is simply true (i.e., the first token corresponds to Zeno's utterance, while the second corresponds to Chrysippus's). But then, from this fact alone, it would seem to follow that—whatever the compositional rules associated with the language are—they are going to have exceptions. For here we have two tokens, made

[12] I should stress that when I say that these languages have 'classical semantic values' I mean that the semantic values of atomic expressions (such as the truth predicate) are always classical. I do *not* mean that the semantic values of *sentences* are always classical: for the semantic value of a sentence is its truth value, and the classical truth values are just t and f; whereas in the sort of languages that I have described some sentences (such as Liar sentences) will be neither true nor false; and these one can very naturally think of as having *non*-classical semantic values.

up of the same words, arranged in the same way, but *differing* in truth value.[13] Further, given that the compositional rules are going to allow exceptions in this way, then one would expect the semantic values associated with T (for example) to be of the familiar classical sort: for if one is going to treat the Liar sentence as an exception to the compositional rules, then there would seem to be little motivation for *also* introducing a new sort of semantic value for T to take.

Thus, formal theories based on the Chrysippus intuition would seem to be the natural place to look for models of the sort of language that I have proposed. The proponents of these theories do not make the point that their theories yield such models (nor do they seem to see their theories as having any particular relevance for anything like the hierarchy-incoherence dilemma). But in the next section I will argue that these theories do in fact yield just the sort of models that we are after.

3. Avoiding the Dilemma: With Details

There are two formal theories that I know of that are based on the Chrysippus intuition: those of Skyrms [1984] and Gaifman [2000]. For the purposes of illustration I will focus on Gaifman's theory; but one could also use Skyrms's theory for these purposes, if one had some independent reason for preferring that.

In this section I will thus state a version of Gaifman's theory, and explain how it yields natural models of the sort that we are after. The theory takes as 'input' a standard, classically interpreted language L.[14] One then adds to L two new 1-place predicate letters, T and F (for 'true' and 'false'). What the theory amounts to is a procedure for assigning a value to each sentence of the extended language; in particular, one assigns each sentence one of t, f or n (for 'neither true nor false').[15] Further, as I will explain, this procedure can

[13] An alternative possibility would be to try to respect the intuition via some sort of context-sensitivity. For example, the first token could be interpreted as on Kripke's approach, while the second could be interpreted as containing some sort of 'higher level' truth predicate. This would not seem to be the most natural approach, however (for it seems that Zeno and Chrysippus mean the same thing by 'true', and similarly for the other words they use). Further, this is not the approach that extant theories based on the intuition take, and so I will ignore the possibility of this approach below.

[14] For simplicity, I will assume that every member of the domain of L is denoted by some closed term of L.

[15] I should say that in stating this version of Gaifman's theory I will make various minor changes to his notation and terminology (in part so as to fit with that used elsewhere in the paper).

also be thought of as simultaneously constructing classical semantic values for T and F: thus, to assign a sentence t is to put it into the extension of T, whereas to assign it f or n is to put it *outside* of this extension; and, similarly, to assign a sentence f is to put it into the extension of F, whereas to assign it t or n is to put it outside of this extension.

I should note that this version of Gaifman's theory is one that he mentions only in passing, and it is quite different from the version that he emphasizes. (Indeed, he probably should not be taken to be committed to the version of the theory stated here.) The version that Gaifman emphasizes is concerned exclusively with truth and falsity for tokens, whereas the version stated here is concerned rather with truth and falsity for sentences (i.e., types rather than tokens). A full treatment of truth (and falsity) must ultimately encompass both sentences and tokens; and a full treatment of the 'Chrysippus' phenomenon in particular must distinguish different tokens of the same sentence (since the utterances of Zeno and Chrysippus above are two tokens of the same sentence). However, for the purposes of this paper it will not be essential to do this, and so I will keep things simple by focusing exclusively on sentences. Further, one will still get instances of the Chrysippus-phenomenon, even within this simplified framework: for example, in the case of a pair of sentences (i.e., types, not tokens) of the form ¬Ta and ¬Tb, where a and b are distinct names, each denoting ¬Ta; in such a case, the formal theory will give the result that ¬Ta is neither true nor false (like Zeno's utterance), while ¬Tb is simply true (like Chrysippus's). Thus, even within this framework, we will get instances of the phenomenon in question.

The procedure for assigning sentences values operates via three rules: one for assigning the 'standard' values t and f, and two 'failure' rules for assigning n. The first of these rules (the 'standard values rule') is straightforward, and as follows. At a typical point in the procedure, one has assigned some sentences a value, while others are yet to receive one. Thus, one has already have partially constructed the final semantic values of T and F: for one has determined of the t-sentences that they will end up in the extension of T, and of the f- and n-sentences that they will not; and similarly for the extension of F (but with t and f switched). Thus, for many sentences of the language, one will already have done enough to determine what the standard compositional rules will say about them (i.e., what they will say about them, given the final semantic values for T and F,

together, of course, with the semantic values of the expressions of the initial language). For example, if A is Tc, and c denotes a sentence that one has already determined will end up in the extension of T, then one has already done enough to determine that the compositional rules will say that A is true. The *standard values rule* simply says that, in such a case, if a sentence has not already been assigned a value, then one may assign it the value that the compositional rules will say that it should get.[16] It is the antecedent clause of this rule (requiring that the sentence is yet to receive a value) that ultimately allows for exceptions to the standard compositional rules: for in the case of a sentence that has been given the value n, the procedure does *not* allow one to reassign it a standard value via this rule, and so in the final interpreted language such a sentence will be an exception to the compositional rules (because *they* would always deliver t or f).

I now move on to describing the failure rules (i.e., the rules for assigning n to sentences). Thus, on the way of thinking about things proposed in this paper, these are the rules that determine where exactly the exceptions to the compositional rules will occur. The basic idea behind these rules is that one should assign n to a sentence if it could never receive a standard value in the 'proper' way (i.e., via an application of the standard values rule). More specifically, the idea is that these sentences are identifiable by the sort of 'referential networks' they belong to. Thus, in the case of a Liar sentence ¬Tc (where c denotes this very sentence), this sentence belongs to a 'loop': if one tries to work out whether it is true, one is sent to consider the denotation of c (i.e., the sentence itself). Because of this the sentence could never receive a value via the standard values rule: for, before it could, one would have to have determined whether the denotation of c (i.e., the

[16] More precisely, the rule is as follows. A given stage of the evaluation procedure corresponds to a partial function from sentences into {t,f,n}, which in turn corresponds to a 'partial interpretation' of the language. I.e., an interpretation in which T and F are assigned pairs of extensions and anti-extensions: so the extension of T will contain the t-sentences, and the anti-extension will contain the f- and n-sentences, together with everything in the domain that is not a sentence of the language (and similarly for F, but with t and f switched). The standard values rule then says that if a sentence A is 'satisfied' by this partial interpretation, and A is yet to receive a value, then one may assign A t; and, similarly, if ¬A is 'satisfied', and A is yet to receive a value, then one may assign A f. But—of course—there are various ways of understanding 'satisfies' here. The simplest is via the strong Kleene scheme, and so it is this version of the rule that I will use in this statement of the theory. But alternatives in terms of some version of the supervaluationist scheme are also natural here (see §4).

It is important to note that the use of partial interpretations here is simply an instrumental device: it is not that in the final language T or F will be interpreted by pairs of sets; rather, their semantic values will be *single* sets (i.e., extensions). Partial interpretations are simply a natural tool to use in the construction of these extensions (since at a given stage we *have* only *partially* constructed these!).

sentence itself) is in the extension of T; but this will only have been determined once the sentence has been assigned a value. Similar examples of loops are {Ta}, where a names the sentence Ta; {Tb, Fd}, where b denotes Fd and d denotes Tb; and {Te → 0 = 1}, where e denotes Te → 0 = 1. Thus, the first failure rule assigns n to sentences that belong to such loops.[17]

It is this rule (in combination with the standard values rule) that allows the final language to respect the Chrysippus intuition. For suppose that c denotes ¬Tc. Then {¬Tc} will be a loop, and so it will be assigned n via an application of this rule. But now suppose that b is a distinct name of this sentence. Then, once ¬Tc has been assigned n, ¬Tb can be assigned t via the standard values rule (for one has now done enough to determine that the standard compositional rules will say that ¬Tb is true, since in assigning ¬Tc n, one has determined that it will be *outside* the extension of T; further, the standard values rule can be applied to ¬Tb, because *it* will not yet have been assigned a value—which is of course what distinguishes it from ¬Tc, which *has* already been assigned one).

The second failure rule assigns n to sentences that belong to a different sort of referential network: namely, 'groundless' sets. Thus, an example of such a set is {Ta_1, Ta_2, …, Ta_n, …} where, for each i, a_i denotes Ta_{i+1}. As in the case of loops, it is clear that the members of such a set could never receive a standard value in the normal way (i.e., via the standard values rule). Thus, the second failure rule assigns n to the members of such groundless sets; and so, as with the members of loops, these sentences will, in the final language, be exceptions to the compositional rules.

I will describe these rules more precisely in a moment, but I will first say something about the end result. Thus, it will be easy to see that by repeatedly applying these three rules one will eventually be able to assign t, f or n to each sentence of the language.[18] The end result of this process can be naturally thought of as follows. We started with a standard interpreted language L, and we have now constructed classical

[17] I will state the rule more precisely below, but note that it is only sentences that say something *semantic* about themselves that count as belonging to loops (for example). Thus, e.g., even if a denotes a = a, {a = a} will *not* count as a loop in the relevant sense.

[18] Although it is less easy to see, this assignment of values is also *unique* (i.e., it does not depend on the order in which one applies the rules to particular sentences). For a theorem to this effect see Gaifman [2000].

semantic values for the two 1-place predicate letters that were added to it. Specifically, the extension of T is the set of sentences we have assigned t to, while the extension of F is the set of sentences that we have assigned f to. Further, the sentences of our new language are all either true, false or neither (i.e., depending on which they value they received). In the case of sentences that are true or false, it is easy to see that these will have the truth value that is determined by the standard compositional rules: for such a sentence must have been assigned t or f via an application of the standard values rule; and this rule assigns a sentence a value only if it has already been determined that the standard compositional rules will—on the basis of the final extensions of T and F—deliver *that* value for the sentence.[19] On the other hand (as I have noted), in the case of a sentence that is neither true nor false in the final language (such as a Liar sentence), this will be an exception to these standard compositional rules (for these rules always say that a sentence is either true or false). Finally, this language is of course also one that contains its own truth and falsity predicates: for the true sentences of the language are those that were assigned t, and these are precisely the contents of the extension of T (and, similarly, the false sentences of the language are precisely the contents of F). Thus, the language is of just the sort that we hoped for: for it contains its own truth predicate, and it is a language with classical semantic values, but where the compositional rules associated with these have exceptions. We saw above that the general approach proposed here would seem to offer the natural way out of the hierarchy-incoherence dilemma. What we are now seeing is that—contrary to an initially natural worry—this approach can in fact be rigorously developed. It would seem, therefore, that we are making a strong case for the proposed approach.

I must now finish the task of stating the failure rules of the procedure. For this we need the notion of 'calling directly', which is as follows. Thus, let S be some stage in the procedure of assigning values to sentences (i.e., S can be thought of as a partial function from the sentences of the language into {t,f,n}). Say that a sentence A is 'determined' at

[19] More precisely, the point is as follows. The standard values rule assigns t (f) to a sentence A if A (¬A) is satisfied by a given partial interpretation. But the final semantic values *extend* the semantic values of this partial interpretation. Thus it follows that the standard compositional rules, using the final semantic values of T and F, will deliver the same result for A as the standard values rule did. (This last point follows from the 'monotonicity' of the strong Kleene evaluation scheme, together with the fact that the strong Kleene scheme reduces to the standard compositional rules in the case of a 'total' interpretation.)

this stage if one has already done enough to fix which standard value the compositional rules will say that A should take;[20] and say that A is 'undetermined' at S otherwise. A *calls directly* B at S if there are A_1, …, A_n that are undetermined at S such that: $A_1 = A$; for each $i < n$, A_{i+1} is either an immediate sentential component[21] or an instance of A_i; and A_n is of the form Tc or Fc for some closed term c denoting B. For example, it is easy to see that if E is ¬Ta, where a denotes E, and E has not yet been assigned a value at S, then E will call itself directly at S. Similarly, if G is Td → 0 = 1, where d denotes G, and G has not received a value at S, then G will also call itself directly at S. Next, say that A *calls* B at S if there are A_1, …, A_n such that: $A_1 = A$; for each $i < n$, A_i calls directly A_{i+1} at S; and $A_n = B$.

The rules are then stated in terms of this notion of calling. A set of sentences X is a *loop* at S if each member of X calls each member of X at S, and no member of X calls any non-member of X at S. For example, it is easy to see that if E is once again ¬Ta, where a denotes E, and E is yet to receive a value at S, then {E} will be a loop at S; similarly, if G is Td → 0 = 1, where d denotes G, and G has not received a value at S, then {G} will also be a loop at S. The first failure rule (the *loop rule*) is as follows: if X is a loop at S, then one may simultaneously assign n to each member of X. This rule will thus assign n to Liar sentences, truth tellers, the members of 'Liar-cycles', sentences such as Tb → 0 = 1 (where b denotes this sentence), and so on.

The second failure rule, concerned rather with 'groundless' sets, is as follows. A set of sentences X is *groundless* at S if: (i) every member of X calls some member of X at S; (ii) no member of X calls a non-member of X at S; and (iii) no non-empty subset of X is a loop at S. For example, if (for each i) $A_i = Ta_{i+1}$, where a_{i+1} denotes A_{i+1}, and S is undefined on each A_i, then $\{A_1, A_2, \ldots, A_n, \ldots\}$ will be groundless at S. A groundless set at S is *complete* at S if it contains every sentence that calls a member of it at S. The second failure rule (the *groundless sets rule*) is as follows: if X is a complete groundless set at S, then one may simultaneously assign n to each member of X.

---

[20] More precisely, S is determined if either A or ¬A is satisfied by the partial interpretation of our language that corresponds to S (see note 16). As with the standard values rule, a number of different understandings of 'satisfied' are possible here, but in this statement I will stick with the strong Kleene understanding.

[21] Thus, the sole immediate sentential component of ¬E is E, the sole immediate sentential components of E ∧ G are E and G, and so on.

That completes the description of the formal procedure. I hope to have made clear that the end result would seem to be a natural model of the sort of language that we need to understand, if we are to avoid the hierarchy-incoherence dilemma, and the expressive restrictions that it leads to.

In the rest of the paper I will (in §4) consider some objections to the proposed approach, and (in §5) consider what the implications for logic are, if the proposed approach is correct.

4. Objections and Replies

Thus, in this section I will consider two objections to the proposed approach. The first is as follows.

(O1) In criticizing approaches in §1 you focused on general truth predicates. In particular, you bemoaned the 'hierarchy-horn' of the dilemma on the basis that it would prohibit the expression of unrestricted generalizations such as 'Nothing is both true and false'. But does the approach that you have proposed not have a similar consequence? For, perhaps your approach allows for general truth predicates. But even if generalizations like that just mentioned are thus rendered 'expressible', they will still not be *true*. For the formal theory that you stated in §3 will assign $\neg\exists x(Tx \wedge Fx)$ (for example) the value n.[22] —Is that really such an improvement?

This is a completely fair objection to the version of the proposed approach described in the last section; i.e., to that in terms of the models given there. Fortunately, however, there are more refined models that address this objection. Thus, the problem with generalizations that (O1) raises stems from the fact that the models of §3 use Kleene's strong scheme: in particular, from the fact that we used this scheme in deciding when we had done enough to determine what the standard compositional rules will say about a given sentence (see note 16). However, as I noted, although the models that use this scheme are perhaps simplest, there are alternatives that use different ones—specifically,

[22] For, under the strong Kleene scheme, this sentence will be satisfied by a partial interpretation only if every sentence of the language is in either the anti-extension of T, or the anti-extension of F. This sentence will thus be eligible to be assigned t (via the standard values rule) only once *every* sentence has been assigned a value—which of course means that this sentence will *never* be assigned t. (It will also never be assigned f; rather, it will be assigned n.)

that use versions of the supervaluationist scheme.[23] For example, any version of this scheme that restricts attention to total interpretations in which the extensions of T and F must be disjoint (a very natural restriction!) will deliver the result that $\neg\exists x(Tx \wedge Fx)$ will come out as true (i.e., it will be assigned t in the models that use this scheme). Indeed, more generally, the more that one refines the supervaluationist scheme, the more such generalizations will come out as true. Thus, although (O1) does point to a real problem with the models of the last section, it would seem to be one that can be overcome.[24]

The second objection I will consider is as follows.

(O2) In §1 you criticized the approaches considered there for leading to expressive restrictions. But doesn't your approach lead to expressive restrictions too, as follows? Thus, consider a Liar sentence A = ¬Tc (where c denotes this sentence). This sentence is not true (on your approach). Fortunately, you can (truthfully) express this fact (on your approach): using a sentence ¬Tb, where b is a new name of A.

OK. But now consider a 'strengthened' paradox as follows: thus, here d denotes the following sentence (which I will call E); and two sentences count as 'similar' here if you can get from one to the other by substituting coreferential terms.

$\forall x$('x is similar to d' $\rightarrow \neg Tx$).

Now, any sentence 'similar' to E will be untrue on your approach (just like the original Liar sentence was). But *this* fact cannot be expressed using the trick that you used before: because let g be some new name of E, and let E' be the result of substituting g for d in E; then E' will simply be one of the sentences 'similar' to E!

Of course, maybe even in *this* case there is some *other* trick that you can use (e.g., you could use not a new name of E but rather some predicate that applied exclusively to E, or the conjunction of E with itself). But surely at *some* point you will run out of tricks—i.e., surely one will eventually be able to find a more inclusive notion of 'similarity' that will lead to an important fact that one simply *cannot* (on your approach) truthfully express. Meaning that you will have expressive restrictions too![25]

---

[23] A partial interpretation P satisfies a sentence A under the basic supervaluationist scheme iff every 'total' interpretation extending P satisfies A. Alternative versions of the scheme restrict attention only to those total interpretations satisfying some further condition.

[24] The contrast with, for example, Kripke's approach should be noted. For that approach also works with a number of different evaluation schemes (including versions of the supervaluationist scheme). However, moving to these schemes does *not*, in this case, help with the problem: for it is easy to see that exclusion negation will still lead to the hierarchy-incoherence dilemma (the argument of §1 goes through essentially unchanged).

[25] A similar objection could be raised in terms of sentences that say of members of groundless sets that they are untrue. For, in the models presented in §3, such sentences will themselves be neither true nor false; and this would thus constitute another sort of expressive restriction faced by the proposed approach (at least as developed in §3). This version of the objection could be responded to similarly to the way in which I respond to (O2) in the text. The reason I focus on (O2), rather than on this alternative objection, is that there are versions of the proposed approach in which some sentences that say of members of groundless sets that they are untrue are themselves *true*. (E.g., versions of the proposed approach based on versions of

In fact, I am inclined to give the objector the benefit of the doubt here. That is, it seems at least plausible that, given enough tinkering, one will eventually be able to come up with some notion of 'similarity' that will do what she wants it to: i.e., yield a sentence E* along the lines of her E, but such that there is no natural way to express the fact that all sentences 'similar' to E* are untrue (because any sentence expressing this fact would itself be 'similar' to E*, and so would say of *itself* that it is untrue). And—if such a notion of 'similarity' *can* be found—there would indeed seem to be an expressive restriction that the proposed approach leads to.

However, this expressive restriction would seem to be fundamentally different from those discussed in §1. For, consider again Kripke's approach, for example. In that case either one cannot express the general notion of truth (if one takes the hierarchy-horn), or one cannot express some extremely simple, and apparently perfectly coherent, logical notion (i.e., exclusion negation, if one takes the incoherence-horn). Either way, the expressive restrictions entailed will prohibit natural claims that have nothing in particular to do with paradoxes (for example, 'Everything is either true or not', using a general notion of truth and the natural but prohibited form of negation). This would seem a much more problematic restriction than that raised by (O2). In the latter case, all we have are certain 'strong' paradoxes that we cannot say certain things about without *ourselves* saying something paradoxical (i.e., something which is neither true nor false). This is a real expressive restriction. But not being able to say certain things about certain paradoxical sentences seems far less objectionable than not being able to express important general claims that do not seem to have anything in particular to do with paradoxes. Thus, (O2) does not establish a problem with the proposed approach comparable to those raised in §1.

5. Implications for Logic

In this final section I will consider the implications for logic, if the approach that I have proposed is correct.

Gaifman's theory with this feature: see Gaifman [2000: 113–18].) In contrast, the expressive restrictions gestured at in (O2) do not seem avoidable.

Thus, the idea behind this approach is that there are exceptions to the compositional rules associated with a language. But we have also seen—in effect—that on the natural development of this approach, the principles of classical logic also admit of exceptions. For, if ¬Tc is a Liar sentence (i.e., if c denotes this sentence), then ¬Tc will be neither true nor false. But if b is a distinct name of this sentence, then ¬Tb will simply be true. And, similarly, b = c will also of course be true (since b and c denote the same sentence; it is only sentences containing semantic predicates such as T or F that will ever fail to be true or false on the proposed approach). This means that an instance of the following classically valid argument-scheme will *not* on this approach be truth preserving (here A(x) is a formula, and d and e are closed terms):

A(d), d = e |= A(e).

That is, on the proposed approach, there are exceptions (in this sense) to this classically valid argument-scheme.

Further, other examples are easy to come by. To illustrate, consider again the models of §3.[26] Thus, consider again a Liar sentence ¬Tc (call this B). Once this sentence has been assigned n via the loop rule, 0 = 0 ∧ B (for example) will be assigned t via the standard values rule. Conjunction elimination (in particular, P ∧ Q |= Q) will thus also have exceptions on this approach. Conjunction introduction will as well: for let a denote the sentence 0 = 0 ∧ ¬Ta (for example; call this sentence E). Then {E} will be a loop, and so it will be assigned n. But once it has been, ¬Ta (i.e., the second conjunct of E) will be assigned t (as, of course, will 0 = 0 be). Each conjunct will thus be true, even though the conjunction is not (giving an exception to P, Q |= P ∧ Q). Therefore, just as there are exceptions to the standard compositional rules, so there are exceptions to almost any logical rule one can think of.[27]

---

[26] That is, I will once again for the purposes of illustration use the formal theory stated in §3 (i.e., the version of the theory of Gaifman [2000] stated there). However, it is plausible that any natural development of the basic idea behind the proposed approach will share the features I am about to describe, at least in essentials. In particular, everything I will say about the models of §3 will apply to the alternatives that use a version of the supervaluationist scheme. (Gaifman, in part because he does not emphasize the version of his theory stated in §3, does not point out that his theory is 'logically exceptionalist' in the sense discussed in the text; but it would seem to be one of the most interesting and important aspects of this theory.)

[27] 'Almost' because there will be some rules without exceptions: e.g., P |= P and P, ¬P |= Q (and |= P ∨ ¬P on supervaluationist versions of the theory).

Just as the proposed approach is to be sharply distinguished from those on which the truth predicate receives a new sort of semantic value, so it is to be distinguished from those on which a new logic is proposed.[28] For it is not that, on the proposed approach, although classical logical rules have exceptions, there is some alternative logic whose rules will always preserve truth. Rather, as I have said, almost any logical rule one can think of will have exceptions on this approach.[29] The situation is thus as follows. Classical logic is correct in the sense that if $A_1, \ldots, A_n \models A_{n+1}$ is classically valid, and A_{n+1} receives a standard value, *then*: if $A_1, \ldots, A_n$ are all true, so is A_{n+1}. But in cases in which the conclusion fails to be true or false, even classically valid argument-schemes can have exceptions—just as (on the proposed approach) the classical compositional rules will have exceptions. There is not some alternative logic whose rules do not have this feature. Rather, on the proposed approach, logical rules, like compositional ones, admit of exceptions.

Indeed, it is hardly surprising that, having set out to avoid the hierarchy-incoherence dilemma, we have ended up with an approach that does not propose a new logic. For the most natural way of generating a new logic is by introducing some new (i.e., non-classical) sort of semantic value. But we saw that any approach that does this would seem to face the hierarchy-incoherence dilemma. Thus, we instead developed an approach with familiar classical semantic values, but where the compositional rules associated with these have exceptions. It was to be expected that this would result in 'classical logic with exceptions', rather than an alternative logic with exceptionless rules.

I hope, then, to have gone some way towards developing an approach to truth and the Liar paradox that avoids the hierarchy-incoherence dilemma, and the expressive restrictions that it leads to.[30]

---

[28] Approaches on which a new logic is proposed include the version of Kripke's approach discussed in §1 (i.e., the strong Kleene version), and the approaches of Maudlin [2004] and Field [2008].

[29] Thus, although rules of certain special sorts—e.g., those in which the conclusion is among the premises, or in which the premises can never all be true—will not have exceptions on this approach (see note 27), the system comprised of only these rules appears too minimal to count as an 'alternative logic'.

[30] I would like to thank Brad Armour-Garb, Andrew Bacon, George Bealer, Jc Beall, David Chalmers, Adam Elga, Hartry Field, Eric Guindon, Elizabeth Harman, John Morrison, Agustín Rayo, Zoltán Gendler Szabó, and members of a seminar at Yale for comments and discussion.

References

Field, H. 2008. *Saving Truth from Paradox*. Oxford: Oxford University Press.

Gaifman, H. 2000. Pointers to Propositions. In A. Chapuis and A. Gupta (eds), *Circularity, Definition, and Truth*: 79–121.

Gupta, A. 1982. Truth and Paradox. *Journal of Philosophical Logic* 11: 1–60.

Gupta, A. and N. Belnap. 1993. *The Revision Theory of Truth*. Cambridge, MA: MIT Press.

Herzberger, H. G. 1982. Notes on Naive Semantics. *Journal of Philosophical Logic* 11: 61–102.

Kripke, S. 1975. Outline of a Theory of Truth. *Journal of Philosophy* 72: 690–716.

Maudlin, T. 2004. *Truth and Paradox: Solving the Riddles*. Oxford: Clarendon Press.

Skyrms, B. 1984. Intensional Aspects of Self-Reference. In R. L. Martin (ed.), *Recent Essays on Truth and the Liar Paradox*: 119–31. Oxford: Clarendon Press.