The Semantic Paradoxes and the Paradoxes of Vagueness

Hartry Field*

March 30, 2003

Both in dealing with the semantic paradoxes and in dealing with vagueness and indeterminacy, there is some temptation to weaken classical logic: in particular, to restrict the law of excluded middle. The reasons for doing this are somewhat different in the two cases. In the case of the semantic paradoxes, a weakening of classical logic (presumably involving a restriction of excluded middle) is required if we are to preserve the naive theory of truth without inconsistency. In the case of vagueness and indeterminacy, there is no worry about inconsistency; but a central intuition is that we must reject the factual status of certain sentences, and it hard to see how we can do that while claiming that the law of excluded middle applies to those sentences. So despite the different routes, we have a similar conclusion in the two cases.

There is also some temptation to connect up the two cases, by viewing the semantic paradoxes as due to something akin to vagueness or indeterminacy in semantic concepts like 'true'. The thought is that the notion of truth is introduced by a schema that might initially appear to settle its extension uniquely: the schema

(T) $\mathsf{True}(\langle p \rangle)$ if and only if p,

(where 'p' is to be replaced by a sentence and '$\langle p \rangle$' by a structural-descriptive name of that sentence). But in fact, this schema settles the extension uniquely only as applied to "grounded" sentences; whether a given "ungrounded" sentence is in the extension of 'true' will be either underdetermined or overdetermined (determined in contrary ways). And this looks rather like what happens in cases of vagueness and indeterminacy: our practices with a vague term like 'bald', or a term like 'heaviness' (which in the mouths of many is indeterminate between standing for mass and standing for weight), don't appear to uniquely settle the reference or extension of the term. (There isn't such a clear distinction between underdetermination and overdetermination in these cases: e.g., it may seem underdetermined whether Harry is bald in that Harry isn't a paradigm case of either baldness or non-baldness, but it may seem overdetermined in that Sorites reasoning may be used to argue that he is bald and also to argue that he is not bald. Similarly, a theorist who makes no distinction between mass and weight hasn't decided which one 'heaviness' stands for, so it may seem underdetermined; but he may attribute to "heaviness" both features true only of mass and features true only of weight, so it may seem overdetermined.)

*New York University. Email: hf18@nyu.edu

In this paper I will argue that an adequate treatment of each of the two phenomena (the semantic paradoxes and vagueness/indeterminacy) requires a nonclassical logic with certain features—features that are roughly the same for each of the two phenomena. This suggests that there might be a common logical framework, and I will propose a framework that seems adequate for treating both. The core logic for the unified framework is sketched in Section 5 (which is rather technical), though much of the discussion before and after (e.g. the discussion of rejection in Section 3 and the discussion of defectiveness in later sections) is highly relevant to the unified treatment. Sections 1, 2, 7 and 8 are primarily concerned with the semantic paradoxes, and Sections 3, 4, 6 and 9 deal primarily with vagueness and indeterminacy (especially in the case of 3 and 4), but there are close interconnections throughout.

1 The semantic paradoxes and attempts to resolve them in classical logic

In this section and the next I will discuss the naive theory of truth and some of the semantic paradoxes which threaten to undermine it. As I've noted, the naive theory of truth includes all instances of the schema (T) above. Perhaps more centrally, it includes the principle that '$\mathsf{True}(\langle p \rangle)$' and '$p$' are always intersubstitutable. (Of course I'm restricting to languages without quotation marks, intentional contexts, and so forth; also without ambiguity, and where there are no relevant shifts of context. Although it may be unnecessary, we can exclude denotationless terms as well.) In classical logic, the schema implies the intersubstitutivity and conversely; whether either direction of the implication holds in a non-classical logic depends on the details of that logic, though I think that there is reason to prefer logics that are "classical enough" for both directions of the implication to hold.

In the context of a very minimal syntactic theory that allows for self-reference, the naive theory of truth is inconsistent in classical logic: we can construct a sentence Q_0 ("the Liar") that is provably equivalent to $\neg\mathsf{True}(\langle Q_0 \rangle)$, so in classical logic we can derive the negation of an instance of (T). It is totally unpromising to blame the problem on the syntactic theory: among other reasons, there are very similar paradoxes in the naive theory of satisfaction which don't require syntactic premises. The real choice is, do we restrict classical logic or restrict the truth schema (and hence its classical equivalent, the intersubstitutivity of '$\mathsf{True}(\langle p \rangle)$' with '$p$').

The attractions of keeping classical logic sacrosanct are powerful, so let's look first at the prospects of a satisfactory weakening of the naive theory of truth within classical logic. By a satisfactory weakening, I mean one that serves the purposes that the notion of truth is supposed to serve, e.g. as a device of making and using generalizations that would be difficult or impossible to make without it. I think it is pretty clear that any weakening of the naive theory of truth will *adversely affect* the ability of 'True' to serve these purposes: the intersubstitutivity of '$\mathsf{True}(\langle p \rangle)$' with '$p$' is very central to the purposes the truth predicate serves. Still, if there were a sufficiently powerful but consistent classical-logic substitute for the naive theory, we might be able to learn to live with it.

In fact, though, I think that restoring consistency requires massive revisions in ordinary principles about truth, revisions that would be very hard to live with. I won't try to fully establish this here, but will make a few observations that give some evidence for it. To begin with an obvious point, the problem in classical logic isn't simply that we can't assert all instances of schema (T), it is that there are instances (such as the one involving Q_0) that we can *dis*prove; and in classical logic, that's equivalent to proving the disjunction of

(1) $p \wedge \neg \mathsf{True}(\langle p \rangle)$

and

(2) $\mathsf{True}(\langle p \rangle) \wedge \neg p$

for certain specific p. Now, it would seem manifestly unsatisfactory to have a theory that proves an instance of (1): how can we assert p and then in the same breath assert that what we've just asserted isn't true? But it also seems manifestly unsatisfactory to have a theory that proves an instance of (2): once we've asserted $\mathsf{True}(\langle p \rangle)$, we're surely licensed to conclude p, so going on to assert $\neg p$ just seems inconsistent. We seem to have, then, that it would be manifestly absurd to have a theory that either proves an instance of (1) or proves an instance of (2). Of course, a classical theory doesn't have to do either: it must prove the disjunction (given that it meets the minimal requirements on allowing self-reference)[1], but it can remain silent on which disjunct to assert. But remaining silent doesn't seem a satisfactory way to resolve a problem: if you have committed yourself to a disjunction of thoroughly unsatisfactory alternatives, it would seem you're already in trouble, even if you refuse to settle on which of these unsatisfactory alternatives to embrace.[2]

I will not further discuss the option of biting the bullet in favor of (2), or the option of accepting the disjunction of (1) and (2) while remaining artfully silent about which disjunct to accept.[3] But I'll say a bit more about the option of biting the bullet in favor of (1).

A superficially appealing way to bite the bullet for option (1) is to say that schema (T) should be weakened to the following:

($T^\star$) If $\mathsf{True}(\langle p \rangle)$ or $\mathsf{True}(\langle \neg p \rangle)$, then $\mathsf{True}(\langle p \rangle)$ if and only if p.

(Proponents of this often introduce the term 'expresses a proposition', and say that $\langle p \rangle$'s expressing a proposition suffices for the consequent of ($T^\star$) to hold and that the antecedent of ($T^\star$) suffices for $\langle p \rangle$ to express a proposition.) It is easily seen that ($T^\star$) is equivalent to the simpler schema

[1] As mentioned above, we wouldn't need even these minimal requirements if we focused on satisfaction rather than truth, and used the heterologicality paradox.

[2] Note that the situation is far worse than for supervaluationist accounts of vagueness. Such accounts allow commitment to disjunctions where we think it would be a mistake to commit to either disjunct. But there the only problem with choosing one disjunct over the other is that the choice seems quite arbitrary; the disjuncts are not thoroughly unacceptable, as they seem to be in the case of the paradoxes.

[3] My own dissatisfaction with the "artful silence" option is not entirely due to the general consideration just raised, but also to the fact that a consistent view of this type must exclude so many natural principles. See [11] and [18] for some important limitations on such theories.

($T^{\star\star}$) If $\mathsf{True}(\langle p\rangle)$ then p.[4]

Obviously these equivalent schemas can't be anything like complete theories of truth: for they are compatible with nothing being true, or with no sentence that begins with the letter 'B' being true, or innumerably many similar absurdities. To get a satisfactory theory of truth that included them, one would have to add a substantial body of partial converses of ($T^{\star\star}$); and one would presumably also want principles such as

(TPMP) $\mathsf{True}(\langle p\rangle) \wedge \mathsf{True}(\langle p \supset q\rangle) \supset \mathsf{True}(\langle q\rangle)$,

or better, the generalized form of this (that whenever a conditional and its antecedent are true, so is the consequent: that is, that modus ponens is truth-preserving).[5] But whatever the details of the supplementation, theories based on ($T^{\star}$) or its equivalent ($T^{\star\star}$) are prima facie unappealing because they require a great many instances of (1). Obviously the Liar sentence Q_0 is one example: since we can prove that $Q_0 \equiv \neg\mathsf{True}(\langle Q_0\rangle)$, ($T^{\star\star}$) yields both Q_0 and $\neg\mathsf{True}(\langle Q_0\rangle)$. But in addition, Montague [19] pointed out that ($T^{\star\star}$) plus (TPMP) plus the very minimal assumption that all theorems of quantification theory are true yields a proof of the untruth of some instances of ($T^{\star\star}$): that is, there is a sentence M (a slight variant of Q_0) such that we can prove

$\neg\mathsf{True}(\langle$ If $\mathsf{True}(\langle M\rangle)$ then $M\rangle)$.[6]

It seems highly unsatisfactory to put forward a theory of truth that includes ($T^{\star\star}$), and use it to conclude that some instances of ($T^{\star\star}$) (including specific instances that you can identify) aren't true.

---

[4] ($T^{\star}$) *implies* ($T^{\star\star}$)*:* Suppose $\mathsf{True}(\langle p\rangle)$; then by ($T^{\star}$), $\mathsf{True}(\langle p\rangle) \equiv p$, which with $\mathsf{True}(\langle p\rangle)$ yields p; so we have $\mathsf{True}(\langle p\rangle) \supset p$. ($T^{\star\star}$) *implies* ($T^{\star}$)*:* This requires two instances of ($T^{\star\star}$), both (i) $\mathsf{True}(\langle p\rangle) \supset p$ and (ii) $\mathsf{True}(\langle\neg p\rangle) \supset \neg p$. Suppose $\mathsf{True}(\langle p\rangle)$; then by (i), p; so $\mathsf{True}(\langle p\rangle) \equiv p$. Alternatively, suppose $\mathsf{True}(\langle\neg p\rangle)$; then by (ii), $\neg p$, and by (i), $\neg\mathsf{True}(\langle p\rangle)$; so again $\mathsf{True}(\langle p\rangle) \equiv p$. So $\mathsf{True}(\langle p\rangle) \vee \mathsf{True}(\langle\neg p\rangle) \supset (\mathsf{True}(\langle p\rangle) \equiv p)$.

[5] Equivalently (given a minimal assumption about how elimination double-negation leaves truth unaffected),

$\mathsf{True}(\langle p \vee q\rangle) \wedge \mathsf{False}(\langle p\rangle) \supset \mathsf{True}(\langle q\rangle)$,

or the generalized form of that; where '$\mathsf{False}(\langle p\rangle)$' means '$\mathsf{True}(\langle\neg p\rangle)$'.

[6] Let R be the conjunction of the axioms of Robinson arithmetic (which is adequate to construct self-referential sentences). Standard techniques of self-reference allow the construction of a sentence N that is provably equivalent, in Robinson arithmetic, to $\mathsf{True}(\langle R \supset \neg N\rangle)$; M will be $R \supset \neg N$. Since this is provable in Robinson arithmetic, then $R \supset [N \supset \mathsf{True}(\langle R \supset \neg N\rangle)]$ is a theorem of quantification theory, hence so is its quantificational consequence

$[\mathsf{True}(\langle R \supset \neg N\rangle) \supset (R \supset \neg N)] \supset (R \supset \neg N)$;

so the claim that that is True is part of the truth theory, and that together with (TPMP) yields

$\mathsf{True}[\langle\mathsf{True}(\langle R \supset \neg N\rangle) \supset (R \supset \neg N)\rangle] \supset \mathsf{True}(\langle R \supset \neg N\rangle)$.

So *if our truth theory proves the negation of the consequent*, it proves the negation of the antecedent, which is the desired negation of the attribution of truth to an instance of ($T^{\star\star}$).

It remains only to show that a proof theory with ($T^{\star\star}$) and arithmetic does prove $\neg\mathsf{True}(\langle R \supset \neg N\rangle)$, but that's easy: from ($T^{\star\star}$) we get $\mathsf{True}(\langle R \supset \neg N\rangle) \supset (R \supset \neg N)$, which given arithmetic yields $\mathsf{True}(\langle R \supset \neg N\rangle) \supset \neg N$; but since N is provably equivalent to $\mathsf{True}(\langle R \supset \neg N\rangle)$, this yields $\neg\mathsf{True}(\langle R \supset \neg N\rangle)$.

It is sometimes thought that one can improve this situation by postulating a hierarchy of ever more inclusive truth predicates True_σ, and for each one adopting $(T^{\star\star}{}_\sigma)$, i.e. the analog of $(T^{\star\star})$ but with 'True_σ' in place of 'True'. (The subscripts are notations for ordinals; the idea is that there will be truth predicates for each member of an initial segment of the ordinals, with no largest σ for which there is a truth predicate. There is no notion of truth$_\sigma$ for variable σ.) For any ordinal σ for which we have such a predicate, we will be able to derive $\neg\mathsf{True}_\sigma(\langle$ If $\mathsf{True}_\sigma(\langle M_\sigma\rangle)$ then $M_\sigma\rangle)$ for a certain sentence M_σ that contains 'True_σ'; but we will also be able to assert $\mathsf{True}_{\sigma+1}(\langle$ If $\mathsf{True}_\sigma(\langle M_\sigma\rangle)$ then $M_\sigma\rangle)$, and this is sometimes thought to ameliorate the situation.

But even if it does, the cost is high. I can't discuss this fully, but will confine myself to a single example. Suppose I tentatively put forward a "theory of truth"—more accurately, a theory of the various truth$_\sigma$s—that includes all instances of $(T^{\star\star}{}_\sigma)$ for each of the truth predicates, together with general principles such as $(TPMP_\sigma)$ for each of the truth predicates, and various partial consequences of each of the $(T^{\star\star}{}_\sigma)$. Someone then tells me that my theory has an implausible consequence; I can't quite follow all the details of his complicated reasoning, but he's a very competent logician and the general strategy he describes for deducing the implausible consequence seems as if it should work, so I come to think he's probably right. Since the consequence still seems implausible, it is natural to conjecture that my theory of truth is wrong—or at least, to consider the possibility that it is wrong and discuss the consequences of that. It is natural to do this even if I have no idea *where* it might be wrong. But I *can't* conjecture this, or discuss the consequences of it, since I have no sufficiently inclusive truth predicate. ('Wrong' means 'not true'.)[7]

And a more specific conjecture, that my theory isn't true$_\sigma$ for some specific σ, won't do the trick. For one thing, *I already know* for each of my truth predicates true$_\sigma$ that not all of the assertions of my theory are true$_\sigma$; after all, it was because I knew that certain instances of $(T^{\star\star}{}_\sigma)$ couldn't be true$_\sigma$ that I was led to introduce the notion of truth$_{\sigma+1}$. Might I get around that problem by finding a way of specifying for each sentence A of my theory of truth a σ_A such that A will be "true$_{\sigma_A}$ if it is true at all" (if you'll pardon the use of an unsubscripted truth predicate)? I doubt that one can find a way to specify such a σ_A for each A: the fact that many principles of a decent truth theory contain quantifiers that range over arbitrary sentences and hence sentences that include arbitrarily high truth$_\sigma$ predicates gives serious reasons for doubt. But even if one can do that—indeed, even if one can specify a function f mapping each sentence A of the theory into the corresponding σ_A—it wouldn't fully get around the problem. For it could well be that for each σ, I would be confident that all members of $\{A|f(A) = \sigma\}$ are true$_\sigma$; a doubt that there is a σ such that not all members of $\{A|f(A) = \sigma\}$ are true$_\sigma$ does not entail a doubt for any specific σ. It is the more general doubt, that there is a σ such that not all members of $\{A|f(A) = \sigma\}$ are true$_\sigma$, that my story motivates; but that more general doubt is unintelligible according to the hierarchical theory, for in treating quantification over the ordinal subscripts as intelligible it violates the principles of the hierarchy.

[7] If the theory were finitely axiomatized I could avoid the use of a truth predicate, but it isn't: that's guaranteed by the need of a separate instance of $(T^{\star}{}^{\star}_\sigma)$ for arbitrarily high σ (or rather, for arbitrarily high σ such that 'true$_\sigma$' is defined).

There is much more that could be said about these matters, but I hope I've said enough to make it attractive to explore an option that weakens classical logic.

2 Semantic paradoxes: a non-classical approach

The most famous non-classical resolution of the paradoxes, due to Kripke [14], employs a logic K_3 that can be read off the strong Kleene truth tables.

More exactly, suppose we assign each sentence A a semantic value $||A||$ of either 1, 0, or $\frac{1}{2}$, with the assignment governed by the following rules:
$||A \wedge B||$ is $min\{||A||, ||B||\}$
$||A \vee B||$ is $max\{||A||, ||B||\}$
$||\neg A||$ is $1 - ||A||$
$||A \supset B||$ is $||\neg A \vee B||$, hence $max\{1 - ||A||, ||B||\}$.
(Think of 1 as the "best" value, 0 as the "worst", and $\frac{1}{2}$ as "intermediate". It may seem more philosophically natural to avoid assigning the value $\frac{1}{2}$, and to instead regard certain sentences as simply having no value assigned to them; but obviously these two styles of formulation are intertranslatable, and the formulation that uses the value $\frac{1}{2}$ allows for a more compact presentation at several points.) Assuming everything to have a name, as I will for simplicity,[8] we also set $||\forall x A|| = min\{||A(x/c)||\}$ and $||\exists\, x A|| = max\{||A(x/c)||\}$. Then Kripke shows that if we start with the language of arithmetic or some other language adequate to syntax, and any arithmetically standard model M for it that is evaluated by these rules, then we can extend the model by designating a subset of M as the extension of 'True' (leaving the ontology and the extension of the other predicates alone) in such a way that for every sentence A, $||\mathsf{True}(\langle A \rangle)||$ will be the same as $||A||$ (and where only objects that satisfy 'Sentence' satisfy 'True'.) More generally, if sentences B and C are alike except that some occurrences of A in one of them are replaced by $\mathsf{True}(\langle A \rangle)$ in the other, then $||B||$ will be the same as $||C||$.

One way to look at this is as showing that we can keep the intersubstitutivity of $\mathsf{True}(\langle A \rangle)$ with A in a revised logic K_3. In K_3 we call an inference *valid* if under every assignment to atomic sentences, if the premises have semantic value 1 then so does the conclusion. And we call a statement *valid* if it has semantic value 1 under every assignment to atomic sentences. Note that instances of the law of excluded middle $(A \vee \neg A)$ comes out invalid: they can have value $\frac{1}{2}$. (Indeed, no statement in this language is valid, though many of the familiar classical rules are valid.) Kripke's result shows that in the logic so obtained,[9] we can consistently assume that for every sentence A and every pair of sentences B and C that are alike except that some occurrences of A in one of them are replaced by $\mathsf{True}(\langle A \rangle)$ in the other, the inference from B to C and from C to B are valid. This is one of the two components of the naive theory of truth, and is not consistently obtainable in classical logic.

[8] Alternatively, we could extend the assignment of semantic values to pairs of formulas and functions assigning objects to variables.

[9] Indeed, even in a slightly expanded logic K_3^+ that includes disjunction elimination as a meta-rule; this rule becomes relevant when one considers adding new validities involving new vocabulary.

It is not entirely clear that this use of nonclassical logic is what Kripke is recommending in his discussion of the strong Kleene version of his theory of truth: some of his remarks suggest it, but others suggest a classical-logic theory later formalized by Feferman ([3], pp. 273-4). The classical-logic Kripke-Feferman theory postulates truth-value gaps: it says of certain sentences, such as the Liar, that they are neither true nor false. ('False' is taken to mean 'has a true negation', so the claim is that neither they nor their negations are true.) As a consistent classical theory, it gives up on the equivalence between '$\mathsf{True}(\langle p \rangle)$' and '$p$'. (The Kripke-Feferman theory is one of the ones that commits itself to disjunct (1) in the previous section.) The Kripke theory in its non-classical version has no commitment to truth-value gaps: indeed, since the whole point of the theory is to maintain the equivalence between '$\mathsf{True}(\langle p \rangle)$' and '$p$', the assertion of $\neg[\mathsf{True}(\langle p \rangle) \vee \neg(\mathsf{True}(\langle \neg p \rangle)]$ would be equivalent to the assertion of $\neg[p \vee \neg p]$; that entails both p and $\neg p$ in the logic, and in this logic as well as in classical that entails everything. So it is very important in the Kripke theory (on its non-classical reading) *not* to commit to truth-value gaps. It will give certain sentences the value $\frac{1}{2}$, but that is not to be read as "neither true nor false".

I think the non-classical reading of Kripke is the more interesting one, and I will confine my discussion to it. I think there are two main problems with the Kripke theory (on this reading).

Perhaps the more serious of the problems is that the logic is simply too weak: as Feferman once remarked, "nothing like sustained ordinary reasoning can be carried out in [the] logic" ([3], p. 264). One symptom of this is that not even the law $A \supset A$ is valid: since $A \supset B$ is equivalent to $\neg A \vee B$, this follows from the invalidity of excluded middle. And note that the intersubstitutivity of $\mathsf{True}(\langle A \rangle)$ with A guarantees that $\mathsf{True}(\langle A \rangle) \supset A$ and its converse are each equivalent to $A \supset A$; since the latter isn't part of the logic in the Kripke theory, neither half of the biconditional $\mathsf{True}(\langle A \rangle) \equiv A$ is validated in Kripke models. So one consequence of the first problem for the Kripke theory is that it does not yield the full naive theory of truth.

The other problem for the Kripke theory, the "revenge problem", has been more widely discussed, but I think *much* of that discussion has been vitiated by a confusion between the non-classical version of Kripke's theory and the classical Kripke-Feferman theory: much of it has been based on falsely supposing that the Kripke theory is committed to truth-value gaps. The only real revenge problem for the non-classical Kripke theory has to do with the fact that the "defectiveness" of sentences like Q_0 is inexpressible in the theory, and there is a worry that if we were to expand the theory to include a "defectiveness predicate" the paradoxes would return. I will be proposing a theory that has much more expressive power than the Kripke theory, and which avoids the revenge problem by having the means to express the defectiveness of paradoxical sentences like Q_0 without this leading to inconsistency.

Returning to the first of the two problems, a natural idea for how to avoid it is to add a new conditional $\rightarrow$ to the Kleene logic, which does obey the law $A \rightarrow A$. There have been many proposals about how to do this; unfortunately, most of them do not enable one to consistently maintain the intersubstitutivity of $\mathsf{True}(\langle A \rangle)$ with A (or even the truth schema $\mathsf{True}(\langle A \rangle) \leftrightarrow A$ which that implies

given the law $A \rightarrow A$). In fact, I know of only two workable proposals for how to do this, both by myself; and one of them ([6]) is not very attractive. (There is also a proposal in Brady [1], which is not an extension of Kleene logic but only of a weaker logic FDE, which is a basic relevance logic.) These theories not only contain $A \rightarrow A$, they also contain a substitutivity rule that allows the inference from $A \leftrightarrow B$ to $C \leftrightarrow D$ when C and D are alike except that one contains B in some places where the other contains A; thus the logic is "classical enough" for the two components of the classical theory of truth to be equivalent.

I will say a little bit about the more attractive of the two theories ([9]). As with Kripke's construction, we start out with a base language that doesn't include 'True', or the new '$\rightarrow$', and with a classical model for this base language whose arithmetical part is standard. The semantics of the theory—which I'll call the Restricted Semantics, since I will generalize it in Section 5—is given by a transfinite sequence of Kripke-constructions. At each stage of the transfinite sequence ("maxi-stage"), we begin with a certain assignment of values in $\{0, \frac{1}{2}, 1\}$ to sentences whose main connective is the '$\rightarrow$'. Given such an assignment of values to the conditionals, Kripke's method of obtaining a minimal fixed point enables us (in a sequence of "mini-stages within the maxi-stage") to obtain a value for every sentence of the language, in such a way as to respect the Kleene valuation rules and the principle that $\mathsf{True}(\langle A \rangle)$ always has the same value as A. It remains only to say how the assignment of values to conditionals that starts each maxi-stage is determined. At the 0^{th} stage it's simple: we just give each conditional value $\frac{1}{2}$. At each successor stage, we let $A \rightarrow B$ have value 1 if the value of A at the prior stage is less than or equal to the value of B; otherwise we give it the value 0. At limit stages, we see if there is a point prior to the limit such that after that point (and before the limit), the value of A is always less than or equal to that of B; if so, $A \rightarrow B$ gets value 1 at the limit. Similarly, if there is a point prior to the limit such that after that point (and before the limit), the value of A is always greater than that of B, then $A \rightarrow B$ gets value 0 at the limit. And if neither condition obtains, $A \rightarrow B$ gets value $\frac{1}{2}$ at the limit. That completes the specification of how each maxi-stage begins; to repeat, it serves as the input to a Kripke construction that yields values at that stage for every sentence.[10]

In typical cases of sentences that are paradoxical on other theories, the values oscillate wildly from one (maxi-)stage to the next. But we can define the "ultimate value" of a sentence to be 1 if there is a stage past which it is always 1; 0 if there is a stage past which it is always 0; and otherwise $\frac{1}{2}$. It turns out that there are ordinals Δ ("acceptable ordinals") such that for any nonzero β, the value of every sentence at stage $\Delta \cdot \beta$ is the same as its ultimate value. (This is the "Fundamental Theorem" of [9].) Since the Kleene valuation rules are satisfied at each stage, this shows (among other important things) that the ultimate values obey the Kleene rules for connectives other than $\rightarrow$. As remarked, this construction validates naive truth theory, both in truth schema and intersubstitutivity form. (It validates it in a strong sense: it not only shows naive truth theory to be consistent, it shows it to be "consistent with any arithmetically

[10] Obviously there is a similarity to the revision theory of Gupta and Belnap [12]; but they use a revision rule for the truth predicate instead of for the conditional, and get a classical logic theory (one of the ones that refuses to commit between (1) and (2)).

standard starting model"—*conservative*, in one sense of that phrase. For a fuller discussion see [9], note 27.)

One question that arises is the relation between the $\rightarrow$ and the $\supset$. $\rightarrow$ is not truth functional,[11] but one can construct a table of the possible ultimate values of $A \rightarrow B$ given the ultimate values of A and of B:

$A \rightarrow B$	$B = 1$	$B = \frac{1}{2}$	$B = 0$
$A = 1$	1	$\frac{1}{2}$,0	0
$A = \frac{1}{2}$	1	1,$\frac{1}{2}$	$\frac{1}{2}$,0
$A = 0$	1	1	1

It is evident from this table that $A \rightarrow B$ is in some ways weaker and in some ways stronger than $A \supset B$. However, from the assumption that excluded middle holds for A and for B, we can derive $(A \supset B) \leftrightarrow (A \rightarrow B)$ (and $(A \supset B) \equiv (A \rightarrow B)$). Moreover, from the assumption that excluded middle holds for each atomic predicate in a set, we get full classical logic for all sentences built up out of just those predicates. Thus the logic is a generalization of classical, and reduces to classical when appropriate instances of excluded middle are assumed. One way to look at the matter is that the logic without excluded middle is the basic logic, but in domains like number theory or set theory or physics where we want excluded middle, we can simply assume all the instances of it in that domain as non-logical premises; this will make the logic of those domains effectively classical. It is only for truth and related notions that we get into obvious trouble from assuming excluded middle: there excluded middle gives inconsistency, given the naive theory of truth.

I think this is a much more attractive resolution of the paradoxes than any of the classical ones. One of its most attractive features has to do with a widely held view that any resolution of the paradoxes simply breeds new paradoxes: "revenge problems". I claim that there are no revenge problems in this logic. More particularly, you can state in this logic the way in which certain sentences of the logic are "defective"; because you can do so, and because there is a consistency proof of naive truth theory in the logic, the notion (or notions) of defectiveness cannot generate any new paradoxes. I will discuss this in Sections 7 and 8.

I will make one remark now, which is that like the non-classical version of the Kripke theory, this is not a theory that posits truth-value gaps. In particular, we can't assert of the Liar sentence that it isn't either true or false. Nor can we assert that it *is* either true or false. Situations like this, where we can't assert either a claim or its negation, may seem superficially like the situation that I complained about in the case of certain classical resolutions of the paradox, where we are committed to a disjunction in which each disjunct has bad consequences, but try to avoid those bad consequence by refusing to decide which of the two disjuncts to assert. But in fact the nonclassical situation isn't like that at all. It is true that in the nonclassical examples we would have a problem if we asserted A and we would have a problem if we asserted $\neg A$ (where A is a classically paradoxical sentence). But what made that so problematic in the classical case was that there we were committed to the claim $A \vee \neg A$. We're

[11] At least not in these values; but see [8] or [7] for an enriched set of semantic values in which it is.

not committed to that in the non-classical case, so our refusal to commit to either the classically paradoxical A or to its negation is not a defect in the account. Similarly, we're not committed to the claim that either A lacks truth value or it doesn't lack truth value, so the refusal to commit to A's being "gappy" or to its being "non-gappy" is no defect.

3 Vagueness and indeterminacy

Before discussing the revenge problem, let's move away from the semantic paradoxes to other quasi-paradoxes.

Many members of the right-to-life movement think that there is a precise nanosecond in which a given life begins, though we may not know when it is. Most of us think that this view is absurd, but Timothy Williamson [23] has in effect offered an interesting argument that the right-to-lifers are correct on this point. The initial argument goes as follows.

Select a precise moment about a year before Jerry Falwell's birth, and call it 'Time 0'. For any natural number N, let 'Time N' mean 'N nanoseconds after Time 0'. By the law of excluded middle, we get each instance of the following schema:

> (3P) (Falwell's life had begun by time N) $\vee\neg$(Falwell's life had begun by time N).

From a finite number of these plus the fact that Falwell's life hadn't begun by time 0 plus the fact that it had begun by time 10^{18}, plus the fact that for any N and M with $N < M$, if Falwell's life had begun by time N then it had begun by time M, a minimal amount of arithmetic and logic yields that

> (F) There is a unique N_0 such that Falwell's life had begun by time N_0 and not by time $N_0 - 1$.

But then it seems that there is a fact of the matter as to which nanosecond his life began, viz. that between time $N_0 - 1$ and time N_0 (inclusive of the latter bound but not the former). That is the initial argument. And the most obvious way around it is to question the use of excluded middle.

There have, of course, been attempts to get around the right-to-lifer's conclusion without giving up classical logic: e.g. by introducing a notion of determinate truth and determinate falsehood such that sentences of form 'Falwell's life began in the interval $(N - 1, N]$' are neither determinately true nor determinately false. But Williamson has given extensions of the initial argument that close off most of these attempts: the basic strategy is to argue that even if such sentences are conceded to be neither determinately true nor determinately false, in whatever sense of determinateness one favors, it's hard to see why this should give any sense of non-factuality to the question of when his life began, given the commitment to (F). Even if I concede that there's no "determinate" truth here, in whatever sense I may give that phrase, why can't I wonder what the unique N_0 is, or wonder whether it is even or odd? Why can't I be very worried about the possibility that the unique N_0 occurred before I performed a certain act, or very much hope that N_0 is odd? And even if I take the question

of whether it is odd to be beyond the scope of human knowledge, why can't I imagine an omniscient god who (by hypothesis of his omniscience) knows the answer; or a Martian who, though not knowing everything, knows this? And so forth. But if I do wonder these things or have worries or hopes like this or concede the possibility of beings with such knowledge, all pretense that I am regarding the question as non-factual seems hollow. In the past I've tried to find a way around this kind of argument, in part by a nonstandard theory of propositional attitudes within classical logic, but I've come to see this task as pretty hopeless. It now seems to me that rejecting some of the instances (3P) of excluded middle is the only viable option (short of giving in to the right-to-lifers on this issue).

But will the no-excluded-middle option work any better? Let's first get clear on an issue (which could have been raised in connection with the semantic paradoxes too) of what it is to "reject" certain instances of excluded middle. We don't reject all of them, only some; what exactly is this difference in attitude we have between those that we reject and those that we don't?

First of all, "reject" can't mean "deny", that is, "assert the negation of". Suppose we deny an instance of (3P), that is, assert

> (3N) $\neg$[(Falwell's life had begun by time N) $\vee\neg$(Falwell's life had begun by time N)].

The expression in brackets is a disjunction, and surely on any reasonable logic a disjunction is weaker than either of its disjuncts. So denying the disjunction has got to entail denying each disjunct, and so asserting (3N) clearly commits us to asserting both of the following:

> (4a) $\neg$(Falwell's life had begun by time N)
>
> (4b) $\neg\neg$(Falwell's life had begun by time N).

But (4b) is the negation of (4a), so (3N) has led to a classical contradiction. And as noted before, the conjunction of a sentence with its negation is also a contradiction in the Kleene logic K_3 described previously, in the sense that there too it implies everything.

Now, that isn't the end of the matter: instead of using K_3 we could follow Graham Priest [20] and opt for a "paraconsistent logic" on which classical contradictions don't entail everything, and therefore aren't so bad as in classical logic. I wouldn't dismiss that view out of hand. But there are problems with using it in the present context. For one thing, since the paraconsistentist accepts (4a), and (3P) is a disjunction with (4a) as one disjunct, the paraconsistentist will accept (3P) as well as (3N): (3P) follows from (4a) on any reasonable logic, including all the standard paraconsistent logics. But then we can argue from (3P) to Williamson's conclusion that there is a unique nanosecond in which Falwell's life began, in precisely the same way as before, so the conclusion has not been blocked. The conclusion has been *denied*—-from (3N) we can conclude that there is not a unique nanosecond during which his life began[12] —-but it has also been asserted. This classical inconsistency is not in itself a problem, it is

[12]Indeed, we can conclude both (i) that there are multiple nanoseconds during which his life began, rather than one, and (ii) that there is no nanosecond during which his life began.

just a further instance of paraconsistentist doctrine that classical inconsistency is no defect; but it is disappointing that we are left in a position of thinking that the right-to-lifers are no less correct to assert that there is a fact of the matter as to the nanosecond in which Falwell was born than we are to deny that there is a fact of the matter.

So rejection must be interpreted in some other way than as denial. A common claim is that to reject A is to regard it as *not true*. The problem with this is that on the most straightforward reading of 'true'—-and the one I took great pains to maintain in the earlier sections on the semantic paradoxes—-the claim that A is true is equivalent to A itself; so asserting that A is not true is equivalent to asserting $\neg A$, and this account of rejection reduces to the previous one.

Perhaps rejection is just non-acceptance? No, that's far too weak. Compare my attitude toward

> (5) Falwell's life began in an even-numbered nanosecond

with my attitude toward

> (6) Attila's maternal grandmother weighed less than 125 pounds on the day she died.

(5) seems intuitively "non-factual", and I reject both it and its negation in the strongest terms. That is not at all the case with (6): I have no reason to doubt that this question is perfectly factual. I don't accept (6) or its negation, for lack of evidence; but I don't reject them either, for given the "factuality" of (6) and its negation I could only reject one by accepting the other. Rejection is more than mere non-acceptance.[13]

The same point arises for the acceptance and rejection of instances of excluded middle. The point would be easier to illustrate here with examples that have less contextual variation than does 'life', and where the higher order indeterminacy is less prevalent; but let's stick to the life case anyway. Suppose I am certain that on my concept of life, if it is determinate that a person's conception occurred during a certain minute then it is indeterminate whether their life began during that minute, but determinate that their life didn't begin before that minute. Then if I knew enough about Falwell to be sure that his conception occurred during some particular precisely delimited minute, and N were the nanosecond marking the end of that minute, then I would reject the corresponding instance of (3P). If however I have no very clear idea how old Falwell is, so that for all I know nanosecond N might be before his conception or after his birth, I will be uncertain about the corresponding instance of (3P):

[13] Rejecting A is also not to be identified with believing it *impossible* that one could have enough evidence to accept A. Why not? That depends on the notion of possibility in question. (a) On any interestingly strong notion of possibility, belief in the impossibility of such evidence does not suffice for rejection: there are intuitively factual 'yes or no' questions (e.g. about the precise goings-on in the interior of the Sun or in a black hole or beyond the event horizon) for which there is no *possible* evidence, but because I take them to be factual I could only reject one answer by accepting the other. (b) On a very weak notion of possibility (e.g. bare logical possibility), we have the opposite problem: even for claims that seem "non-factual", like (5), there is a bare logical possibility that there is such a thing as "living force" and that someone will invent a "living force detector" that could be used to ascertain whether the claim is true.

I will neither accept it nor reject it. (The same point can arise even if I do have detailed knowledge of the times of his conception and birth and the various intermediate stages: suppose that I'm undecided whether there is a God who injects vital fluid into each human body at some precise time, but think that if there is no such God then N would correspond to a borderline case of Falwell's life having begun.) So for instances of excluded middle too, we have that rejection is stronger than mere non-acceptance.[14]

Should the failure of all these attempts to explain the notion of rejection required by the opponent of excluded middle lead us to suppose that there is no way to make sense of the no-excluded-middle position? No, for in fact there is an alternative way to explain the concept of rejection (and it doesn't require a prior notion of indeterminacy). The key is to recognize that the refusal to accept all instances of excluded middle forces a revision in our other epistemic attitudes. A standard idealization of the epistemic attitudes of an adherent of classical logic is the Bayesian one, which (in its crudest form at least) involves attributing to each rational agent a degree of belief function that obeys the laws of classical probability; these laws entail that theorems of classical logic get degree of belief 1. Obviously this is inappropriate if rational agents needn't accept all instances of excluded middle. But allowing degrees of belief less than 1 to some instances of excluded middle forces other violations of classical probability theory. In particular, if we keep the laws

$$P(A \vee B) + P(A \wedge B) = P(A) + P(B)$$

and

$$P(A \wedge \neg A) = 0,$$

then we must accept

$$P(A \vee \neg A) = P(A) + P(\neg A).$$

In that case, assigning degree of belief less than 1 to instances of excluded middle requires that we weaken the law

$$(7)\ P(A) + P(\neg A) = 1$$

to

$$(7_w)\ P(A) + P(\neg A) \leq 1.$$

The relevance of this to acceptance and rejection is that accepting A seems intimately related to having a high degree of belief in it; say, a degree of belief at or over a certain threshold $T > \frac{1}{2}$.[15] So let us think of rejection as the dual

[14] The point arises as well in connection with potentially "ungrounded" sentences that may not be actually "ungrounded". If sentence A is of form "No sentence written in location D is true", and I know that exactly one sentence is written in location D but am unsure whether it is '1+1=3' or A itself, then I am not in a position to *accept or reject* either the sentence A or the sentence $A \vee \neg A$.

[15] We can take T to be 1, but only if we are very generous about attributing degree of belief 1. If (as I prefer) we take T to be less than 1, some would argue that the lottery paradox prevents a strict identification of acceptance with degree of belief over the threshold; I doubt that it does, but to avoid having to argue the matter I have avoided any claim of strict identification.

notion: it is related in the same way to having a low degree of belief, one at or lower than the co-threshold $1 - T$. In the context of classical probability theory where (7) is assumed, this just amounts to acceptance of the negation. But with (7) replaced by (7_w), rejection in this sense is weaker than acceptance of the negation. (It is still stronger than failure to accept: sentences believed to degrees between $1 - T$ and T will be neither accepted nor rejected). I take it that in a case where a sentence A is clearly indeterminate (e.g. case (5), for anyone certain that there is no such thing as "vital fluid"), the degree of belief in A and in $\neg A$ should both be 0.

Some may feel it more natural to say that the degree of belief in "Falwell's life began in an even-numbered nanosecond" should be not the single point 0, but the closed interval [0,1]. That view is easy to accommodate: represent the degree of belief in A not by the point $P(A)$, but by the closed interval $R(A) =_{df} [P(A), 1 - P(\neg A)]$.[16] This is merely a matter of terminology: the functions P and R are interdefinable, and it is a matter of taste which one is taken to represent "degrees of belief". (In terms of R, acceptance and rejection of A go by the lower bound of $R(A)$.)

The value of introducing probabilistic notions is that they give us a natural way to represent the gradations in attitudes that people can have about the "factuality" of certain questions—-at least, they do when higher order indeterminacy is not at issue. To regard the question of whether A is the case as "certainly factual" is for the following equivalent conditions on one's degree of belief to obtain:

$P(A) + P(\neg A) = 1$;
$P(A \vee \neg A) = 1$;
$R(A)$ is point-valued;
$R(A \vee \neg A) = \{1\}$.

To regard it as "certainly nonfactual" is for the following equivalent conditions to hold:

$P(A) + P(\neg A) = 0$;
$P(A \vee \neg A) = 0$;
$R(A) = [0, 1]$;
$R(A \vee \neg A) = [0, 1]$.

In general, the degree to which one believes A determinate is represented (in the P-formulation) by $P(A) + P(\neg A)$; i.e. $P(A \vee \neg A)$; i.e. $1 - w$, where w is the breadth of the interval $R(A)$; i.e. the lower bound of $R(A \vee \neg A)$. In the P-formulation, belief revision on empirical evidence goes just as on the classical theory, by conditionalizing; this allows the "degree of certainty of the determinacy of A" to go up or down with evidence (as long as it isn't 1 or 0 to start with).

The idea can be used not only for examples like the Falwell example, but for potentially paradoxical sentences as well. Consider a sentence S that says that no sentence written in a certain location is true, and suppose that we know that exactly one sentence is written in that location; our degree of belief that the sentence in that location is '$2 + 2 = 4$' is p, our degree of belief that it is '$2 + 2 = 5$' is q, and our degree of belief that the sentence written there is S

[16] Note that $R(A \vee \neg A)$ will always be an interval with upper bound 1; its lower bound will be $1 - w$, where w is the width of the interval $R(A)$.

itself is $1-p-q$, which I'll call r. I submit that our degree of belief in S should be q, our degree of belief in $\neg S$ should be p, and our degree of belief in $S \vee \neg S$ should be $p+q$, i.e. $1-r$. The key point in motivating this assignment is that relative to the assumption that the sentence written there is S, then S and $\neg S$ each imply the contradiction $S \wedge \neg S$, and so $S \vee \neg S$ implies this contradiction as well; given this, it seems clear that if we were certain that the sentence written there were S, then we should have degree of belief 0 in S, in $\neg S$ and in $S \vee \neg S$. As we increase our degree of certainty that the sentence written there is S, our tendency to reject the three sentences S, $\neg S$ and $S \vee \neg S$ should become stronger.

Of course, the idea that we can attribute to an agent a *determinate* P-function (or R-function) is a considerable idealization. Even in the case of classical P-functions, where we don't allow $P(A) + P(\neg A)$ to be less than 1, the issue of whether a person's degree of belief is greater than say 0.7 often seems indeterminate. How are we to make sense of the indeterminacy here? It should be no surprise that on my view, we make sense of this by giving up the presupposition of excluded middle for certain claims of form "X's probability function P_X is such that $P_X(A) > 0.7$".[17] (It isn't that we need to develop the theory of probability itself in a non-classical language; where excluded middle is to be questioned, rather, is in the attribution of a given perfectly classical probability function to a given agent X. If you like, this gives failures of excluded middle for "claims about P_X", though not for claims about individual probability functions specified independently of the agent X.) We can take the same position for non-classical P-functions or R-functions too. I'm inclined to think that there is a strong connection between this indeterminacy in the degree of belief function and "higher order indeterminacy": in cases where X attributes higher order indeterminacy to A, some assertions about the value of $P_X(A \vee \neg A)$ (or $R_X(A \vee \neg A)$) will be ones for which excluded middle can't be assumed. In any case, the indeterminacy in attributions of probability doesn't essentially change the picture offered in the preceding paragraph: to whatever extent that we can say that X's degree of belief function attributes value 1 to $A \vee \neg A$, to precisely that extent we can say that X regards A as certainly factual.

So far I have not said anything about introducing a notion of determinacy into the language. I have argued that *even without doing so*, we can represent a "dispute about the factuality of A" as a disagreement in attitude: a disagreement about what sort of degrees of belief to adopt. An "advocate of the factuality of A" will have a cognitive state in which $P(A \vee \neg A)$ is high (i.e. in which $R(A)$ is close to point-valued). An "opponent of the factuality of A" will have a cognitive state in which $P(A \vee \neg A)$ is low (i.e. $R(A)$ occupies most of the unit interval). I think it important to see that we can do all this *without* bringing the notion of determinacy into the language: it makes clear that there is more substance to a dispute about factuality than a mere debate about how a term like 'factual' or 'determinate' is to be used.

[17] A common suggestion ([17]) is that we should represent the epistemic state of an agent X not by a single probability function but by a non-empty set Σ_X of them. That is in some ways a step in the right direction, but it too involves unwanted precision; and while that could be somewhat ameliorated by going to nonempty sets of nonempty sets of probability functions, or iterating this even further, I think that ultimately there is no satisfactory resolution short of recognizing that excluded middle fails for some attributions.

Still, what we have so far falls short of what we might desire, in that so far we have no means to literally assert the nonfactuality of the question of whether A: having a low degree of belief in $A \vee \neg A$ is a way of *rejecting* the factuality of A, but not of *denying* it. It would be very awkward if we couldn't do better than this: debates about the factuality of questions would be crippled were we unable to treat the claim of determinacy or factuality as itself propositional. What we need, then, is an operator G, such that $\mathsf{G}A$ means intuitively that A is a determinate (or factual) claim, i.e. that the question of whether A is the case is a determinate (or factual) question. Actually it's simpler to take as basic an operator D, where $\mathsf{D}A$ means that it is determinately the case that A. The claim $\mathsf{G}A$ (that it is determinate *whether* A) is the claim that $\mathsf{D}A \vee \mathsf{D}\neg A$. The point of the operator is that though $\neg(A \vee \neg A)$ is a contradiction, $\neg(\mathsf{D}A \vee \mathsf{D}\neg A)$ is not to be contradictory.

"Determinately operators" are more familiar in the context of attempts to treat vagueness and indeterminacy within classical logic, and their use there in representing nonfactuality is subject to a persuasive criticism. The criticism is that whatever meaning one gives to the D operator, it is hard to see how $\neg(\mathsf{D}A \vee \mathsf{D}\neg A)$ can represent the nonfactuality of the question of whether A: for any claims to nonfactuality are undermined by the acceptance of $A \vee \neg A$. But when we have given up the acceptance of $A \vee \neg A$, the criticism doesn't apply.

People can have degrees of belief about determinateness, so their degree of belief function should extend to the language containing D. If we had only first order indeterminacy to worry about, and could stick to the idealization of a determinate degree of belief function P_X for our agent X, some constraints on how P_X extends to the D-language would be obvious: since $P_X(A) + P_X(\neg A)$ represents the degree to which the agent regards A factual, which is $P_X(\mathsf{D}A \vee \mathsf{D}\neg A)$, which in turn is $P_X(\mathsf{D}A) + P_X(\mathsf{D}\neg A)$, we must suppose that $P_X(\mathsf{D}A) = P_X(A)$ for any A.[18] That is, we must regard the lower bound of $R_X(\mathsf{D}A)$ as the same as the lower bound of $R_X(A)$. Indeed, with higher order indeterminacy excluded, $R_X(\mathsf{D}A)$ should just be a point: $P_X(\neg\mathsf{D}A)$ will be simply $1 - P_X(A)$. So the fact that $\mathsf{D}A$ is strictly stronger than A comes out in that $P_X(\neg\mathsf{D}A)$ can be greater than $P_X(\neg A)$ but not less than it, i.e. in that the upper bound of $R_X(\mathsf{D}A)$ can be lower than that of $R_X(A)$ but not greater than it. It is however important to allow for higher order indeterminacy, and there may be some question how best to do so. A proper representation of higher order indeterminacy presumably should allow excluded middle to fail for sentences of form $\mathsf{D}A$, so we want to allow that $P_X(\mathsf{D}A) + P_X(\neg\mathsf{D}A)$ falls short of 1, i.e. that $R_X(\mathsf{D}A)$ not be point-valued. I'm inclined to think that we ought to keep the demand that the lower bound of $R_X(\mathsf{D}A)$ is always the same as that of $R_X(A)$; this would leave the upper bound unfixed.[19] (As noted before, the situation

[18] More generally, we could argue that for any A and B, $P(\mathsf{D}A \wedge B) = P(A \wedge B)$.

[19] There are intuitions that go contrary to this: sometimes we seem prepared to assert A but not to assert "It is determinately the case that A". I'm somewhat inclined to think that this is so only in examples where A contains terms that are context-dependent as well as indeterminate, and that it is so because we give to 'determinately A' a meaning like 'under all reasonable contextual alterations of the use of these terms, A would come out true'; and this seems to me a use of 'determinately' different from the one primarily relevant to the theory of indeterminacy. But I confess to a lack of complete certainty on these points; for instance, another possibility would be to allow that in some contexts the upper bound of $R(A)$ plays a role in governing the assertion of A. (I'd like to thank Richard Dietz, Stephen

is complicated by the fact that there is indeterminacy in the attribution of P-functions or R-functions to the agent, so we can't assume excluded middle for all claims about P_X and R_X. I don't believe that this requires modification of what I've said, but the matter deserves more thought than I have been able to give it.)

I think that what I've said clarifies important aspects of the conceptual role of the determinately operator. Indeed, until recently I thought that not a whole lot more could be said to clarify that operator. I now realize that that former opinion was far too pessimistic: in fact, I'm tempted to say that the determinately operator can be defined in terms of a more basic operator, a conditional '$\rightarrow$'. As we'll see in the next section, such a conditional plays a very central role in the theory of vagueness; in Section 5 I will then make a case that the conditional required is "essentially the same as" the one used in connection with the semantic paradoxes. Starting in Section 6, I will explain how the conditional can be used to define a very good candidate for the determinately operator. The definition of 'determinately' in terms of the conditional would not make the probabilistic laws governing the determinately operator irrelevant: for they would indirectly constrain our degrees of belief in sentences involving the conditional. The conditional, however, can be given a very rich set of deductive relationships; it is these deductive relationships on it, together with the probabilistic constraints on the determinately operator defined from it, together with that definition, which would jointly clarify the conditional and the determinately operator together.

The picture just sketched may eventually need to be complicated slightly: while we can certainly define a *very good candidate for* the intuitive determinately operator D in terms of the conditional, the defined operator D *may* not match the intuitive operator in every respect. But even if this turns out to be so, the intuitive operator and the defined operator will share enough properties for D to provide a very good model of D; in particular, the intuitive laws governing D, including the laws of its interaction with the conditional, will be very close to the laws provable for D. In short: I'm tempted by the view that we have a strict definition of D in terms of '$\rightarrow$', but even if not, what we do have will give rich structural connections between the two that, when combined with the probabilistic account above, do a great deal to settle the meaning of D.

4 The conditional again

What kind of logic do we want to use for vagueness? For reasons mentioned already, it should not include the law of excluded middle. I will now assume that the logic should be like the Kleene logic K_3 (or K_3^+: see note 9) as regards the basic connectives $\neg$, $\wedge$, and $\vee$. This assumption is both plausible and widely accepted, and seems to have the best hope of providing a unification of the logic of vagueness with the appropriate logic for the semantic paradoxes.

Schiffer and Timothy Williamson for discussions of the complications of any extension of talk of probability to a language with a determinately operator, when higher order vagueness is allowed. In particular, Dietz pointed out a substantial problem in an earlier attempt of mine, in a classical-logic context.)

The standard semantics for the logic K_3 (or indeed, K_3^+) is the strong Kleene valuation tables, mentioned previously, on which each sentence receives one and only one of the values 1, 0 and $\frac{1}{2}$. These should not be read 'true', 'false' and 'neither true nor false': then assigning value $\frac{1}{2}$ would be postulating a truth value gap, which we cannot do for reasons given in Section 2. A somewhat better reading would be 'determinately true', 'determinately false', and 'neither determinately true nor determinately false'. That isn't quite right either: among other things, since it is assumed that each sentence has one of the three values, this reading would commit us to the claim that excluded middle holds for attributions of determinate truth and determinate falsehood,[20] thereby ruling out higher order indeterminacy. (This argument assumes that the semantics is given in a classical metalanguage; an alternative is to keep to the readings 'determinately true', 'determinately false' and 'neither of those', but refrain from assuming of every sentence that it has one of the three values. More on this later.) In Sections 5 and 8 I will say a bit more about these issues of how to interpret the semantics. But for now, the readings 'determinately true', 'determinately false', and 'neither determinately true nor determinately false' will be close enough to serve our purposes.

In the non-classical treatment of the semantic paradoxes in Section 2, we saw that there was a strong need to supplement the Kleene logic with a new conditional: a conditional $A \rightarrow B$ not defined in the classical way, as $\neg A \vee B$. For whatever the merits of that definition in a context where we have excluded middle, it fails miserably when excluded middle is abandoned: it doesn't even obey such elementary laws as $A \rightarrow A$, and because of this it is very hard to reason with.

Indeed, the failure of the connective $\neg A \vee B$ (which I'll call the Kleene conditional, and abbreviate $A \supset B$) as a definition of $A \rightarrow B$ is perhaps even more striking in connection with vagueness than it is in connection with the semantic paradoxes. For consider the following claims:

A: There are nearly 1000 red balls in Urn 1

B: There are nearly 1000 black balls in Urn 1.

In fact, let us suppose, there are exactly 947 red balls and exactly 953 black balls in Urn 1. Let us imagine a context where both 947 and 953 seem to be in the borderline region of 'nearly 1000', so that we'd be inclined to regard A and B as each having value $\frac{1}{2}$. But since the number of black balls is closer to 1000 than is the number of red balls, we'd surely want to give "If there are nearly 1000 red balls in urn 1 then there are nearly 1000 black balls in urn 1" semantic value 1. But if we use the connective '$\supset$' to define 'if...then', we don't get this result: '$A \supset B$' will come out with value $\frac{1}{2}$. So we very much need another conditional.

The point is not a new one: nearly all advocates of non-classical logics of vagueness have proposed using a conditional other than '$\supset$'. The most popular expansion of Kleene-logic is Łukasiewicz continuum-valued logic, *aka* "fuzzy

[20] The commitment to one of the three values is $\mathtt{D}T \vee \mathtt{D}F \vee (\neg \mathtt{D}T \wedge \neg \mathtt{D}F)$, which is equivalent to $(\mathtt{D}T \vee \mathtt{D}F \vee \neg \mathtt{D}T) \wedge (\mathtt{D}T \vee \mathtt{D}F \vee \neg \mathtt{D}F)$; since $\mathtt{D}F$ entails $\neg \mathtt{D}T$ and $\mathtt{D}T$ entails $\neg \mathtt{D}F$, this entails $(\mathtt{D}T \vee \neg \mathtt{D}T) \wedge (\mathtt{D}F \vee \neg \mathtt{D}F)$.

logic". For the semantics of this conditional we need to further partition the values other than 0 and 1: instead of just $\frac{1}{2}$, we allow arbitrary real numbers between 0 and 1. The value of $A \wedge B$ is then the minimum of the values of A and B, while the value of $A \vee B$ is the maximum; and the value of $\neg A$ is 1 minus the value of A. Clearly, as far as these connectives go the valuation rules are a generalization of the Kleene rules: the Kleene rules result by restricting attention to the three values 0, $\frac{1}{2}$, and 1. Moreover, taking 1 as the sole designated value, the class of valid inferences in these connectives is unaffected. The point of the new semantics (as far as logic as opposed to pragmatics goes) is that it enables us to give rules for a new connective $\rightarrow$; the valuation rule is that the value of $A \rightarrow B$ is 1 if the value of A is less than or equal to that of B, and otherwise is 1 minus the extent by which the value of A exceeds that of B.

The Łukasiewicz semantics works *reasonably* well for dealing with vagueness, but I have never seen a compelling argument for it. And it does have some intuitive defects: for instance, the linear ordering of values is unintuitive, and leads to the claim that sentences such as

> It is either the case that if Tim is thin then John is old, or that if John is old then Tim is thin

are logical truths. Moreover, using the Łukasiewicz logic would doom all hope of a combined treatment of vagueness and the semantic paradoxes which preserves the naive theory of truth, for the naive theory cannot be preserved in the Łukasiewicz logic ([22], [13]). But despite these doubts about the Łukasiewicz logic, we do need a new conditional. (Whether the one proposed in connection with the semantic paradoxes would do for this purpose is something I will consider in the next section.)

The issue of the conditional is also relevant to an influential argument of Kit Fine's [10]. Fine argued that classical logic accounts of vagueness are far superior to nonclassical logic accounts based on the Kleene semantics, in that the Kleene-based accounts cannot handle "penumbral connections" between distinct vague terms. Fine's point is this. Suppose that the claim that an object b is red and the claim that b is small both get value $\frac{1}{2}$. It is unsurprising that the conjunction and disjunction of the two claims should get value $\frac{1}{2}$; and that is what the Kleene tables say. But how about the claim that b is red and the claim that b is pink, when b is a borderline case of each? Fine thinks that their conjunction ought to get value 0, and that if b is clearly in the red-to-pink region their disjunction should get value 1 even though the disjuncts get the value $\frac{1}{2}$.

I don't find these intuitions as compelling as Fine does, but there is a more neutral way to put his point: we ought to be able to say, somehow, that 'red' and 'pink' are contraries; and we ought to be able to say, somehow, that an object is red-to-pink, using only the terms 'red', 'pink' and logical devices. But the obvious proposals for how to do these things won't work. E.g. we can't say that something is red-to-pink by saying that it is red or pink, since that gets value $\frac{1}{2}$ when the object is on the border between red and pink; and no other logical function in Kleene logic will do any better. Put this way, Fine's objection against the use of the unadorned Kleene logic is compelling.

But once we add a new conditional to the logic, it is much less obvious that there is any problem with penumbral connections. For now we can easily ex-

plain the idea that 'red' and 'pink' are contraries: it consists in the fact that $\square\forall x[x$ is red $\rightarrow x$ is not pink $]$,[21] where $\square$ indicates some kind of conceptual necessity; the claim is that what follows the $\square$ is guaranteed by a conceptual constraint on the simultaneous values that are allowed for 'red(x)' and 'pink(x)'.[22] And we can explain the idea of an object being in the red-to-pink region: "red-to-pink(x)" just means "x is not red $\rightarrow$ x is pink".[23] We can, if we like, even introduce a "pseudo-disjunction" $\sqcup$ and "pseudo-conjunction" $\sqcap$:

$A \sqcup B$ iff $(\neg A) \rightarrow B$

$A \sqcap B$ iff $\neg(A \rightarrow \neg B)$.

So to call something red-to-pink(x) is to say that it is red $\sqcup$ pink, and the contrariness of the two consists in the fact that nothing can be red $\sqcap$ pink. The logic of $\sqcap$ and $\sqcup$ will be slightly odd (just what it is will depend on the logic of $\rightarrow$, obviously),[24] but perhaps it is close enough to the ordinary $\wedge$ and $\vee$ to explain whatever intuitive force there is in Fine's own way of presenting the penumbral connection problem.

As we'll see in Section 6, an appropriate conditional can also be used in clarifying the notion of determinateness, and the related notion of defectiveness. But just what sort of conditional is appropriate?

5 Generalizing the previous conditional

Before deciding what the semantics of a conditional appropriate to vagueness and indeterminacy should be, I need to say something about the role I expect the semantics to play. There are two approaches to giving a semantic account of a language with vague terms.

[21] If $\rightarrow$ were not contraposable, we'd have to add $\square\forall x[x$ is pink $\rightarrow x$ is not red $]$; but both the Łukasiewicz conditional and the conditional introduced in Section 2 are contraposable (as will be the more general conditional introduced in Section 5).

[22] In the case of the Łukasiewicz semantics, the appropriate constraint would be that the values assigned to 'red(o)' and to 'pink(o)' never add to more than 1. In the case of the Restricted Semantics I suggested for the paradoxes, the appropriate constraint would have to be that the values assigned to 'red(o)' and to 'pink(o)' at any sufficiently large stage never add to more than 1. The more general semantics to be introduced in the next section will drop the applicability of stages to sentences like 'red(o)' and 'pink(o)', but will contain a more general sort of variation of extension; but the constraint will be analogous, that the values assigned to 'red(o)' and to 'pink(o)' at any "world" (at least, any "world" near "the actual one") never add to more than 1.

[23] In Łukasiewicz semantics, an object will thus satisfy 'red-to-pink' iff the values assigned to 'red(o)' and to 'pink(o)' add to exactly 1; in the semantics suggested for the paradoxes or its generalization in the next section, the condition is that the values assigned to 'red(o)' and to 'pink(o)' add to exactly 1 at any sufficiently large stage, or at any world near the actual one.

[24] It may be useful to display the appropriate "possible value tables" for $\sqcup$ and $\sqcap$ that we get if we use the semantics outlined in Section 2:

$\sqcup$	$B=1$	$B=\frac{1}{2}$	$B=0$
$A=1$	1	1	1
$A=\frac{1}{2}$	1	$1,\frac{1}{2}$	$\frac{1}{2}$,0
$A=0$	1	$\frac{1}{2}$,0	0

$\sqcap$	$B=1$	$B=\frac{1}{2}$	$B=0$
$A=1$	1	$1,\frac{1}{2}$	0
$A=\frac{1}{2}$	$1,\frac{1}{2}$	$\frac{1}{2}$,0	0
$A=0$	0	0	0

One is to use a perfectly precise metalanguage. This approach has one great advantage and one disadvantage. The great advantage is that we are all used to reasoning classically, and there are no real controversies about how to do it; the semantics will thus be easy to use and have a clear and unambiguous content. The disadvantage is that there is no hope of precisely capturing the content of the sentences of a vague language in a precise metalanguage: the precise metalanguage inevitably draws artificial lines. (This is why higher order vagueness looks like such a problem when we think about it in terms of a precise metalanguage.)

The only way one could conceivably give a semantic theory for a language with vague terms that accurately reflects the content of its sentences is to drop the assumption of excluded middle in the metalanguage. The metalanguage could still be like the metalanguage used before, in assigning semantic values in the set $\{0, \frac{1}{2}, 1\}$ to sentences; but we could not assume the law of excluded middle for attributions of such values.

A semantic theory that avoids excluded middle in the metalanguage may be what we should ultimately aspire to, but it does throw away the great advantage of a semantics in a classical metalanguage, and I think that for the here and now it is more useful to give our semantics in a classical metalanguage. A result of doing this is that we cannot expect our semantics to be *totally* faithful to the language it seeks to represent.[25]

What then should our standards on a semantics in a classical metalanguage be? There is a very modest conception of semantics, which Dummett ([2]) calls "semantics as a purely algebraic tool", where the only point of a semantics is to yield an extensionally adequate notion of logical consequence. If that were the limits on our ambition for a semantics, then it is not obvious that the Restricted Semantics offered in Section 2 in connection with the semantic paradoxes couldn't be simply carried over to the case of vagueness and indeterminacy. In describing the construction there, I stipulated that we were to start out with a perfectly precise base language: one where each atomic predicate gets a 2-valued extension that it keeps at each stage of the construction. I then proposed adding the term 'True', and I took its semantics to have two features that might be taken as typical of indeterminacy: first, 'True' gets only a "3-valued extension" ("positive extension", "negative extension" and "remainder") at each stage, and second, this 3-valued extension varies from stage to stage. (Its varying from stage to stage is connected with the higher order indeterminacy of 'True'.) And so a natural thought is that we might simply allow that predicates like 'red' and 'bald' could be treated like 'True': we could simply assign to 'Red' or 'Bald', at each stage α, a 3-valued extension, with this extension allowed to vary from one stage to the next. There is, actually, a technical constraint we'd want to impose on the way that the 3-valued extension can vary, in order to ensure that we could still add 'True' to the language in accord with

[25] For independent reasons, the clarification of the meaning of logical connectives can't proceed wholly by means of a formal semantics: any formal semantics will itself use logical connectives, often the very ones being "explained", and often in ways that would make the "explanations" grossly circular (e.g. "'not A' is true if and only if A is not true", "'A and B' is true if and only if A is true and B is true", etc.). In my view, part of what clarifies all these notions, the conditional of this section included, is its connection to degrees of belief: see the discussion at the end of Section 3.

the naive truth theory. But the required constraint is fairly evident from the proof of the technical result mentioned in Section 2.[26] With this constraint imposed, we would validate the naive theory of truth *even when there are vague predicates in the language*; and do so in such a way that the same logic of $\rightarrow$ that works when we add '**True**' to a precise language works for vague language generally.

This is all fine if we accept the view that the only role of the semantics is to serve as an "algebraic tool" for getting a consequence relation, but I think it is reasonable to demand more. True, it is inevitable that a classical semantics for a nonclassical language won't reflect the semantics totally faithfully, but we'd like it to represent it as faithfully as possible, and the semantics just suggested seems to me to fall short of reasonable expectations. In particular, it isn't at all clear what the assignment of 3-valued extensions to stages "means" in the case of vague predicates. It does seem natural (at least when one models the semantics in a classical metalanguage, which is what's under consideration here) to think of models that have varying 3-valued extensions. What is not so natural, though, is the well-ordering of these 3-valued extensions into stages. We certainly don't have the clear rules constraining the transfinite sequence of 3-valued extensions of ordinary vague predicates that we have in the case of 'True'. We might be tempted to just require that the 3-valued extension stays the same from stage to stage; but if we do that, the model will have no representation at all of higher order vagueness, and that doesn't seem satisfactory.

What I think is needed is a more substantial generalization of the Restricted Semantics used for the paradoxes; I'll call it the Generalized Semantics. A natural idea for a semantics of vagueness in a classical metalanguage is to employ an infinite set W of *classical partial worlds*, or just *worlds* for short, with one of them @ singled out as privileged. (I prefer to think of them not as alternative possibilities, but as alternative methods for assigning semantic values to actual and possible sentences, given the way the world actually is in precise respects; @ represents "the actual assignment".) We need to equip W with a certain structure; I propose that to each $w \in W$ we assign a (possibly empty) directed family $\mathbb{F}_w$ of nonempty subsets of W, which I call *w-neighborhoods*. To say that $\mathbb{F}_w$ is directed means

(*) $(\forall w \in W)(\forall U_1, U_2 \in \mathbb{F}_w)(\exists U_3 \in \mathbb{F}_w)(U_3 \subseteq U_1 \cap U_2)$.

I will add a few further conditions on the $\mathbb{F}_w$ as we proceed. Think of each w-neighborhood as containing the "worlds" that meet a certain standard of closeness to w. I'm allowing that the standards of closeness for different w-neighborhoods may be incomparable, so that two w-neighborhoods for the same w can have members other than w in common without one being a subset of the

[26] A sufficient constraint on the valuation of the atomic predicates other than '**True**', for a consistent addition of '**True**' to be possible, is that for each n-ary predicate p and each n-tuple of objects o_1,...,o_n, the sequence of members of $\{1, \frac{1}{2}, 0\}$ assigned to $p(o_1, ..., o_n)$ satisfies the following two conditions: (i) it *eventually cycles*, in the sense that there are α and β, the latter non-zero, such that for any two stages $\gamma_1, \gamma_2 \geq \alpha$, the values of $p(o_1, ..., o_n)$ at these stages is the same if γ_1 and γ_2 differ by a right-multiple of β; (ii) unless the "cycle length" (the smallest such β) is 1, then the value of $p(o_1, ..., o_n)$ is $\frac{1}{2}$ at all sufficiently high right multiples of β. As long as this regularity condition is assumed for the base language, the Fundamental Theorem of [9] goes through with no real change. (And it tells us that the regularity condition holds for 'True' as well.)

other. (If you don't want to allow for incomparability, you could replace (*) with the simpler requirement that the members of $\mathbb{F}_w$ are linearly ordered by $\subseteq$; this would simplify the resulting theory a bit without drastically changing its character.)

It would be natural from the motivation to assume that for any w, w is a member of each U in $\mathbb{F}_w$. But for reasons that *might* only be relevant to the semantic paradoxes, I want to allow for the existence of *abnormal* worlds that do not meet this condition; all I assume in general is that @ is normal. I should also say that in my intended applications, $\{@\} \notin \mathbb{F}_@$ (indeed, it should probably be the case that for all w, $\{w\} \notin \mathbb{F}_w$); so each member of $\mathbb{F}_@$ includes a point other than @. The validities that follow do not depend on this, but if $\{@\}$ were a member of $\mathbb{F}_@$ we would get new validities that would rule out higher order indeterminacy and prevent the application of the theory to the semantic paradoxes.

A model for a vague language L will consist of a domain V and for each $w \in W$ an assignment to each n-place atomic predicate of a "3-valued extension" in V^n. (That is, to each $(n+1)$-tuple whose first element is an n-place predicate and whose other elements are in V, we assign at each $w \in W$ a member of $\{0, \frac{1}{2}, 1\}$.) I will later add a constraint on the assignment of these 3-valued extensions to atomic predicates. The valuation rules for the Kleene connectives and quantifiers at a given world are just the usual Kleene rules; no reference to other worlds is required. For '$\rightarrow$' I propose the following:

$|A \rightarrow B|_w$ is
1, if $(\exists U \in \mathbb{F}_w)(\forall u \in U)(|A|_u \leq |B|_u)$;
0, if $(\exists U \in \mathbb{F}_w)(\forall u \in U)(|A|_u > |B|_u)$;
$\frac{1}{2}$ otherwise.

(The stipulation that the members of $\mathbb{F}_w$ are all nonempty is needed to keep the value of $|A \rightarrow B|_w$ unique.) If w is abnormal, $\mathbb{F}_w$ has members that don't contain w; we could without alteration restrict the quantification in the 1 and 0 clauses to such members, given the directedness condition.

I call an inference *universally valid* if in any model and any world w in it, whenever all the premises have value 1 at w, so does the conclusion. I call an inference *strongly valid* if in any model, whenever all the premises have value 1 at all normal worlds, so does the conclusion. And I call an inference *valid* if in any model, whenever all the premises have value 1 at @, so does the conclusion. Validity, strong validity and universal validity for a sentence are just validity, strong validity and universal validity for the 0-premise argument whose conclusion is that sentence. Validity is the notion that will be of direct interest, but information about the other two helps in proving results about validity; besides, it illuminates the semantics to see what is universally valid, what is merely strongly valid, and what is just valid.

The above stipulations effectively generalize what was done before in the case of the semantic paradoxes, though this may not be completely obvious. There, the space W was, in effect, a closed initial segment $[0, \Delta]$ of the ordinals, where Δ was one of the "acceptable ordinals" proved to exist in the Fundamental Theorem of [9];[27] @ was Δ. For any ordinal α in W, the $\mathbb{F}_\alpha$ used in specifying the semantic values of conditionals at α had as its members all (nonempty)

[27] To quell a possible worry, I remark that it is possible to choose the acceptable ordinal "in advance". (Certainly the initial ordinal for any cardinal greater than that of the power set of

intervals of form $[\beta, \alpha)$. So no member of W was normal, not even Δ; however, I used a slightly different definition of validity, so the notion of validity didn't trivialize. How then can I say that the account here "effectively generalizes" the one there? Because the Fundamental Theorem shows that if we redefine $\mathbb{F}_\Delta$ to include intervals of form $[\beta, \Delta]$ (making Δ normal), while leaving the other $\mathbb{F}_\alpha$ as before, then the values of conditionals is unchanged; this allows us to redefine validity in accord with the general account above, and the redefined one and the original are equivalent in extension. In short: the Restricted Semantics sketched in Section 3 could be rewritten so as to be a special case of the one just outlined.

Returning to the general case, let's get some results about which sentences and inferences are valid, strongly valid, and universally valid. Even without imposing any additional constraints, we can easily prove the following:

1. The universal validities include every inference that is valid in the Kleene logic K_3. In addition, the following inferences involving the conditional are universally valid:

$\vdash A \rightarrow A$
$\vdash \neg\neg A \rightarrow A$
$\vdash A \rightarrow A \vee B$ and $\vdash B \rightarrow A \vee B$
$\vdash A \wedge B \rightarrow A$ and $\vdash A \wedge B \rightarrow B$
$\vdash A \wedge (B \vee C) \rightarrow (A \wedge B) \vee (A \wedge C)$
$\vdash (A \rightarrow \neg B) \rightarrow (B \rightarrow \neg A)$
$\vdash (A \rightarrow \neg A) \leftrightarrow \neg(\top \rightarrow A)$
$\vdash \neg(A \rightarrow B) \rightarrow (B \rightarrow A)$
$(A \rightarrow B) \wedge (A \rightarrow C) \vdash A \rightarrow (B \wedge C)$
$(A \rightarrow C) \wedge (B \rightarrow C) \vdash (A \vee B) \rightarrow C$
$\vdash \forall x A \rightarrow At$ (with the usual restrictions on legitimate substitution)
$\vdash \forall x(A \vee Bx) \rightarrow A \vee \forall x Bx$, when x is not free in A

2. In addition, the following are strongly valid:

$A, A \rightarrow B \vdash B$
$\vdash \neg(A \rightarrow B) \rightarrow (A \vee \neg B)$

3. If $A \vdash C$ and $B \vdash C$ are both valid, so is $A \vee B \vdash C$; and analogously for strong validity and universal validity. (This means that we validate K_3^+, not just K_3.)

The proofs of all of 1-3 are completely straightforward; indeed, the only ones that even require the directedness condition are the two universal validities with conjunctions of conditionals as premises.

We can do better if we add two additional constraints on the families $\mathbb{F}_w$. First let's define U_1 is *w-interior to* U_2 to mean: $U_1, U_2 \in \mathbb{F}_w \wedge U_1 \subseteq U_2 \wedge (\forall x \in U_1)(\exists X \in \mathbb{F}_x)(X \subseteq U_2)$. The first constraint is

(i) For any normal world w, every member of $\mathbb{F}_w$ has w-interior subsets.

The second constraint is a generalization of directedness. If $\{$ is a cardinal number, call a family $\mathbb{F}$ of subsets of W $\{$*-directed* if for any collection S of subsets of $\mathbb{F}$ that has cardinality no greater than $\{$, there is a $U \in \mathbb{F}$ such

the domain V will be acceptable; I suspect that there are acceptable ordinals of much lower cardinality.)

that U is contained in each member of S. (Directedness as defined before is 2-directedness; and is equivalent to n-directedness, for any finite $n \geq 2$.) Then letting $||V||$ be the cardinality of the domain V, we stipulate

(ii) For any world w, $\mathbb{F}_w$ is $||V||$-directed.

We can now prove the following additional strong validities:

$A \rightarrow B \vdash (C \rightarrow A) \rightarrow (C \rightarrow B)$

$A \rightarrow B \vdash (B \rightarrow C) \rightarrow (A \rightarrow C)$

$\forall x(Ax \rightarrow Bx) \vdash \forall xAx \rightarrow \forall xBx$

$\forall x(\neg Ax \rightarrow Ax) \vdash \neg\forall xAx \rightarrow \forall xAx$.[28]

Finally, let us now add the previously-promised constraint on the assignment of 3-valued extensions to atomic predicates. Actually I'll impose the constraint originally only for what I'll call "standard" predicates, by which I'll mean, all but special predicates like 'True'. There's no need to be very precise here, because the idea is that for any "nonstandard" predicates we introduce, we'll *prove* that the constraint holds for them too, though this will not be part of the initial stipulation governing their 3-valued extensions. So let p be a standard n-place predicate; we assume

> (**) for each choice of n objects $o_1, ..., o_n$ in V, if $\langle p, o_1, ..., o_n\rangle$ is assigned an integral value (0 or 1) at @, then there is an @-neighborhood throughout which $\langle p, o_1, ..., o_n\rangle$ is assigned that same value.

We can now prove a general lemma: that for any sentence A all of whose predicates are standard,

> If A takes on an integral value at @ then there is a @-neighborhood throughout which A takes on that same value.

The proof is by induction on complexity; the only noteworthy cases are the conditional and the quantifier. For the conditional, the noteworthy fact is that the proof for $A \rightarrow B$ goes not by the induction hypothesis, but rather by (i). (If $|A \rightarrow B|_@ = 1$ then there is an @-neighborhood U throughout which $|A| \leq |B|$; letting U' be @-interior to U, it is clear that $|A \rightarrow B|_p = 1$ for any p in U'. Similarly for 0.) For the quantifier case, we again use (ii).

Given this Lemma, we get two more validities (for sentences all of whose atomic predicates are standard):

[28] Proofs: (I) $A \rightarrow B \vdash (C \rightarrow A) \rightarrow (C \rightarrow B)$: Suppose that w is normal and that $|A \rightarrow B|_w = 1$. Then there is a U in $\mathbb{F}_w$ throughout which $|A| \leq |B|$. By (i), there's a U' that is w-interior to U. Suppose now that $|(C \rightarrow A) \rightarrow (C \rightarrow B)|_w < 1$. Then there's a $p \in U'$ such that $|C \rightarrow A|_p > |C \rightarrow B|_p$. So either $|C \rightarrow A|_p = 1$ or $|C \rightarrow B|_p = 0$. But both lead to contradiction. For instance, in the first case, there's a p-region X throughout which $|C| \leq |A|$; by directedness, there's a p-region $Y \subseteq X \cap U'$; so throughout Y, $|C| \leq |A| \leq |B|$, so $|C \rightarrow B|_p = 1$, giving the desired contradiction. The second case is analogous.

(II) $A \rightarrow B \vdash (B \rightarrow C) \rightarrow (A \rightarrow C)$: Analogous to (I). (Indeed, (II) follows from (I), given the laws already established.)

(III) $\forall x(Ax \rightarrow Bx) \vdash \forall xAx \rightarrow \forall xBx$: Suppose that w is normal and that $|\forall x(Ax \rightarrow Bx)|_w = 1$. Then for all t, $|At \rightarrow Bt|_w = 1$. So for each t, there is a w-region U_t such that $|At| \leq |Bt|$ throughout U_t. By (ii), there is a w-region U such that for each t, $|At| \leq |Bt|$ throughout U, and hence such that $|\forall xAx| \leq |\forall xBx|$ throughout U. (Recall that every element of V is assumed to have a name.) So $|\forall xAx \rightarrow \forall xBx|_w = 1$.

(IV) $\forall x(\neg Ax \rightarrow Ax) \vdash \neg\forall xAx \rightarrow \forall xAx$: Analogous to (III).

$B \vdash A \rightarrow B$
$A, \neg B \vdash \neg(A \rightarrow B)$
The proof of the first is totally trivial (given the Lemma), and the proof of the second almost so (it uses the directedness condition).

The validities here established for the Generalized Semantics include almost all the ones that I derived in [9] for the Restricted Semantics (plus two additional ones that I neglected to consider there); the only one derived there that we don't have here is the relatively minor one $\neg[(C \rightarrow A) \rightarrow (C \rightarrow B)] \vdash \neg(A \rightarrow B)$. Actually we would get that too if we made the additional supposition that there is an @-neighborhood in which all worlds are normal; that assumption seems fairly plausible in applications of the logic to ordinary cases of vagueness, though it does not hold in the Restricted Semantics (where the validity in question holds for a different reason). It seems then that the logic validated by the Generalized Semantics is the main core of what's validated by the Restricted Semantics, and I propose it as the unified logic for vagueness and the paradoxes. (There may be ways to expand it slightly, e.g. to get the above-mentioned law $\neg[(C \rightarrow A) \rightarrow (C \rightarrow B)] \vdash \neg(A \rightarrow B)$.)[29]

One thing that needs to be shown, if this is really to be a successful unification of the logic of vagueness with the logic of the paradoxes, is that the naive theory of truth is not merely "consistent with any *classical* starting model (that is adequate to syntax)", as proved in [9], but "consistent with any *generalized model for the logic of vagueness* (that is adequate to syntax)". More precisely, let L be any language without 'True', but which may contain vague terms and the '$\rightarrow$'; we also suppose that L contains arithmetic, so that syntax can be done within it. Let M be any generalized model of the sort just described for a language without 'True' (where M is built from a set W of worlds with distinguished world @, an assignment to each of an $\mathbb{F}_w$, a domain V, and an assignment of 3-valued extensions in V^n to each n-place predicate that is subject to (**)); the only restriction is that the extension of arithmetical predicates be the same at each world and take on only the values 0 and 1, and that the resulting 2-valued model be a standard model of arithmetic. The claim then is that given any such generalized model M, we can extend it to a model M^* on the same structure that satisfies naive truth theory; by 'extend it' I mean that each predicate other than 'True' has the same extension in M^* as in M. This can in fact be shown, by a straightforward generalization of the proof given in [9]; and (**) turns out to hold for 'True' as well, so that even the final two

[29] One natural expansion–which wouldn't yield that law, but might yield others–is suggested by the fact that in the Restricted Semantics there is a "uniformness" to the structure imposed on the worlds; that is, there is a way to compare the size of α_1-neighborhoods to those of α_2-neighborhoods even when $\alpha_1 \neq \alpha_2$, namely the ordinal that must be added to the lower bound of the interval to get the upper bound. The idea seems natural in the vagueness case too. So in the general case, we might want to add to the structure an equivalence relation $\approx$ on the set $\Re$ of pairs $\langle w, U\rangle$ for which $U \in \mathbb{F}_w$, with $\langle w_1, U_1\rangle \approx \langle w_2, U_2\rangle$ having the intuitive meaning that U_1 is a w_1-neighborhood and U_2 is a w_2-neighborhood and U_1 is like U_2 in size, shape and orientation from its "base point" (w_1 or w_2 as the case may be). Axioms, besides those of being an equivalence relation and $\langle w_1, U_1\rangle \approx \langle w_2, U_2\rangle \supset U_1$ is a w_1-neighborhood and U_2 is a w_2-neighborhood, should certainly include (a) and probably (b):
(a) $\langle w, U_1\rangle \approx \langle w, U_2\rangle \supset U_1 = U_2$
(b) $(\forall w_1, w_2)[\mathbb{F}_{w_1} \neq \varnothing \wedge \mathbb{F}_{w_2} \neq \varnothing \supset (\exists U_1, U_2)[\langle w_1, U_1\rangle \approx \langle w_2, U_2\rangle \wedge$
$(\forall X \subseteq U_1)(\exists Y \subseteq U_2)(\langle w_1, X\rangle \approx \langle w_2, Y\rangle) \wedge (\forall Y \subseteq U_2)(\exists X \subseteq U_1)(\langle w_1, X\rangle \approx \langle w_2, Y\rangle)]$.
Just what additional strength this would bring to the logic I'm not sure.

validities established above hold in the full language.[30] Without going through the details, let me just say that predicates other than 'True' are assigned the same 3-valued extension at every stage of the construction: there is no longer a need to vary them from stage to stage to get reasonable results about higher order vagueness, since that is already handled by the multiple "worlds" in the initial stage. We thus avoid the artificiality that (early in this section) we saw would be required were we to use the Restricted Semantics for vagueness.

6 Defectiveness and determinateness

Whether or not one accepts the unification just proposed between vagueness/indeterminacy and the semantic paradoxes, we've seen that in each case there are strong grounds for giving up the law of excluded middle: for certain sentences A, we should not assert that $A \vee \neg A$. But we should not assert that $\neg(A \vee \neg A)$ either: that lands us in contradiction. How then do we assert the "defectiveness" of A? Not by simply asserting that we are not in a position to assert $A \vee \neg A$: that's too weak, since our inability to assert

> Attila's maternal grandfather was bald on the day he died $\vee\neg$(Attila's maternal grandfather was bald on the day he died)

is due not to conviction that its disjuncts are defective but to ignorance as to whether they are defective. (He might have been clearly bald or clearly non-bald.) Nor can we assert defectiveness by asserting that no possible evidence could lead anyone to assert $A \vee \neg A$; examples about the interiors of black holes in indeterministic universes, or events outside of the event horizon of any regions that can support conscious beings, can easily be constructed to show this. I've suggested that our treating A as certainly defective consists (to a first approximation anyway) in our believing $A \vee \neg A$ to degree 0 (or to degree [0,1], if we use R-degrees instead of P-degrees; that's the same as R-believing A to degree [0,1]). Similarly, our treating A as probably defective consists (to a first approximation) in our believing $A \vee \neg A$ to a low degree. So perhaps we could assert the "defectiveness" of A by saying that we ought to believe $A \vee \neg A$ to degree less than some amount (the amount fixed by the confidence of our assertion of defectiveness).[31] But it's natural to think that we should be able

[30] An alternative way to think of the procedure, which illustrates more clearly the sense in which the proposed account is a unification, is as going from the initial M to a new model M^{**} with the same domain V, but a new set W^* of worlds. On this account, W^* is $W \times [0, \Delta]$, where Δ is a large ordinal (a sufficiently large power of ω), and the new @* is $\langle @, \Delta \rangle$. For $\langle w, \alpha \rangle$ in W^*, $\mathbb{F}_{\langle w,\alpha \rangle}$ consists of precisely the subsets of form $U \times [\beta, \alpha)$ where in the original model $U \in \mathbb{F}_w$ and where $\beta < \alpha$, except in the case where α is Δ, in which case the subsets have form $U \times [\beta, \alpha]$. Predicates other than 'True' get the same value at any $\langle w, \alpha \rangle$ that they got in the original model at w; the value of 'True' at any $\langle w, \alpha \rangle$ is determined by a Kripke construction of the minimal fixed point given the values of the 'True'-free sentences. The model M^* (in the formulation in the main text) is simply the submodel of M^{**} consisting of worlds of form $\langle w, \Delta \rangle$.

The modest additional assumption is that for each $w \in W$, $\mathbb{F}_w$ is $||\Delta||$-directed, not merely $||V||$-directed. This may not really be an additional assumption, as I suspect that we can take Δ to have cardinality $||V||$; but at the moment I don't know how to prove that.

[31] It's worth noting that we don't presently have a satisfactory account of the laws of "probability" (rational degree of belief) for the logic: we do have a satisfactory probability theory

to assert the defectiveness in a more direct way, without bringing ourselves or other believers into the story. Can we do so?

One possibility is simply to introduce an undefined operator $\mathsf{D}A$ into the language, give some laws (e.g. probabilistic laws) for how it works, and define a defectiveness predicate $\mathsf{B}\,\mathsf{A}\,\mathsf{D}(x)$ by $\neg\mathsf{G}(\mathsf{True}(x))$, i.e. $\neg\mathsf{D}(\mathsf{True}(x))\wedge\neg\mathsf{D}\neg\mathsf{True}(x)$. But one worry is that we don't really understand such an operator. In addition, in the semantic paradox case it raises a serious issue of "liar's revenge": there is reason to worry that it might give rise to new paradoxes, e.g. for sentences that assert of themselves that they are not determinately true.

It turns out that in the logic suggested in Section 2 and the generalization of it suggested in Section 5, there is a good case to be made that *the only defectiveness predicates we really need are already definable in the language*, using the new conditional. Obviously that would remove the worry about the intelligibility of the operator, if one grants that the conditional itself is intelligible. Moreover, it would mean that no revenge problem can arise: the construction sketched in Section 2 (and later generalized) shows that we can validate naive truth theory for that language (using a certain non-classical logic), and that means that we can validate it *even for sentences about defectiveness and truth together*, so that there is no revenge problem.

A first stab at defining a determinately operator D might be this:

$$D^?A =_{df} \top \rightarrow A \text{ (or equivalently, } \neg A \rightarrow \bot)$$

where $\top$ is some logical truth (say $B \rightarrow B$, for some arbitrarily chosen B), and $\bot$ is some absurdity. In the "worlds" semantics of Section 5, the value of this at a given world w is 1 if there is a neighborhhod of w throughout which A has value 1; 0 if there is a neighborhhod of w throughout which A has value less than 1 (it needn't be all 0 or all $\frac{1}{2}$, it could be a mixture of the two); and $\frac{1}{2}$ if there is neither of these sorts of neighborhood. The only clear problem with this as a definition of determinately is that if the world w is abnormal, the neighborhood needn't include w, so $D^?A$ could have a higher value than A at w. But the situation can be remedied by a familiar trick: define D as

$$DA =_{df} (\top \rightarrow A) \wedge A \text{ (or equivalently, } (\neg A \rightarrow \bot) \wedge A).$$

(I use a different typeface than for D so as not to prejudge whether D adequately defines the notion we need.)

Inferentially, the proposed definition of D obeys the obvious laws. Because the semantics validates the rule

(8) $B \vdash A \rightarrow B$,

we get the validity of the D-introduction rule

$$A \vdash DA.$$

for the Kleene logic, but it is not obvious how to extend it to the logic that includes the $\rightarrow$. (In what follows I will suggest a possible definition of D in terms of $\rightarrow$; that together with my earlier remark on the desired probability for sentences of form $\mathsf{D}A$ would give a constraint on how the probability theory is extended to the language with $\rightarrow$, but would fall far short of settling it.)

Modus ponens guarantees the validity of $D^?A \vdash A$, but not of $\vdash D^?A \rightarrow A$ (since @ might have abnormal worlds in all its neighborhoods); but now that we've shifted from $D^?$ to D, we get the universal validity of

$\vdash DA \rightarrow A$

(strong D-elimination).

The definition is also quite natural in connection with the Restricted Semantics used for the paradoxes in Section 2: e.g. if A has value less than 1 at a stage, DA has value 0 at the next stage.

Indeed, the merits of this definition of determinateness come out especially clearly in its consequences for the semantic paradoxes, in particular as regards the "revenge problem". It is to this that I now turn.

7 Revenge (1)

A simple illustration of how the determinately operator D works is afforded by the Liar sentence Q_0, which asserts its own untruth. It's clear that on the semantics of Section 2, it must have value $\frac{1}{2}$ at every stage; so the value of DQ_0 and of $D\neg Q_0$ are both 0 at every stage after stage 0, and so $BAD(\langle Q_0 \rangle)$ has ultimate value 1. ('BAD' is defined from D in the same way that 'BAD' is defined from D—see the previous section.)

Once we have a determinately operator D, it's natural to consider a "weakened" liar sentence Q_1, that says of itself that it is not determinately true.[32] But the construction outlined in Section 2 shows that this must be consistently evaluable in the language. Indeed, it isn't hard to see that it's value is $\frac{1}{2}$ at all even ordinals and 1 at all odd ordinals. DQ_1 gets value $\frac{1}{2}$ at all even ordinals and 0 at all odd ordinals; $\neg DQ_1$ thus has the same value as Q_1, as desired. $\neg D\neg Q_1$ gets ultimate value 1, as we might expect: so we can assert that Q_1 is *not* determinately *un*true. As for the claim that Q_1 is determinately *true*, its ultimate value is $\frac{1}{2}$, so we can't assert $DQ_1 \vee \neg DQ_1$ (and indeed, can reject it). So excluded middle can't be assumed (and indeed, can be rejected) *even for claims of determinateness*: that is, we have a kind of second order indeterminacy. But we can assert that Q_1 isn't *determinately* determinately true. So we can assert that Q_1 is BAD_2, where $BAD_2(x)$ means that $\neg DD(\mathsf{True}(x)) \wedge \neg DD\neg\mathsf{True}(x)$. If we rename the earlier predicate BAD as BAD_1, then badness$_1$ entails badness$_2$, but not conversely.

We can now consider a "still weaker" liar sentence, that says of itself that it is not *determinately* determinately true, and so forth. Indeed, we can iterate the determinately operator a fair way through the transfinite: for as long as our system of ordinal notations lasts.[33] For each σ for which there is a notation, we can then consider the σ-Liar Q_σ, which says of itself that it is not D^σ-true (D^σ being the σ-fold iteration of D). We will never be able to assert $D^\sigma Q_\sigma \vee \neg D^\sigma Q_\sigma$, whatever the σ, showing that indeterminacy of arbitrarily

[32] 'Weakened' is in quotes since though Q_1 attributes a weaker property than Q_0 does, it attributes it to a different sentence.

[33] At limit stages we need to form infinite conjunctions; but we can do that, for limits simple enough to have a notation in some reasonable system, by using the truth predicate.

high levels[34] is allowed; but we can assert that Q_σ is not determinately untrue, and not determinately$^{\sigma+1}$ true. (There will be a notation for $\sigma + 1$ whenever there is one for σ.) Defining the predicates BAD_σ in analogy to the above, we get that each is more inclusive than the previous, and for each Liar sentence, there will be a σ for which the Liar sentence can be asserted to be BAD_σ.

This hierarchy of defectiveness predicates has something of the flavor of the hierarchy of truth predicates that we have in the classical case. But it is different in important ways. For one thing, it affects a much more peripheral notion: the relevant notion of defectiveness (or paradoxicalness) plays only a marginal role in the lives of most of us, whereas the notion of truth is ubiquitous. I don't think it is that hard to learn to live with the idea that we do not have a unified notion of defectiveness* that captures all "levels of defectiveness". A second difference is that we can reasonably hope that whatever the σ, our overall theory has the virtue of non-defectiveness$_\sigma$; this is in sharp contrast to the hierarchical classical truth theories, where for each σ there are sentences of our overall theory that we assert while asserting not to be true$_\sigma$. So I don't think that settling for the hierarchy of defectiveness predicates would be debilitating. (Indeed, it is especially un-debilitating if we allow schematic reasoning about levels of determinateness: see [4], [15], and pages 141-3 of [5] for discussions of schematic reasoning in other contexts.)[35]

So far I've been talking about determinately operators and defectiveness predicates that are actually definable in the language. But it might be thought that these definitions do not adequately capture the intuitive notions of determinateness and defectiveness–or to put it in a way that avoids the supposition that there are unique such notions, it might be supposed that there are intuitive notions of determinateness and defectiveness that are not adequately captured by these definitions. (I don't myself know of any clear reason to suppose that there are intuitive notions that the definitions fail to capture, but I see no reason to be dogmatic about the issue.) So we might want to add a new primitive determinately operator, governed by axioms, from which a corresponding defectiveness predicate could be defined. By using the axioms that we know are true of the defined operator D as a guide to the choice of axioms for the primitive operator D, one can have a reasonable hope of achieving a consistent theory of determinateness slightly different from that given by D or any of its iterations. The most natural idea would be to model such a D quite closely on D. In that case, we would expect that iterations of the operator are non-trivial: more particularly, that for each value of σ, the D^σ-Liar can't be asserted not to be D^σ-true but can be asserted not to be $\mathsf{D}^{\sigma+1}$-true.

I don't say that it would be hopeless to add to the language a "maximum strength" determinately operator $\mathsf{D}^\star$ that iterates only trivially, in that $\mathsf{D}^\star\mathsf{D}^\star$ is equivalent to (intersubstitutable with) $\mathsf{D}^\star$; that would give rise to a unified defectiveness predicate $\mathsf{B}\,\mathsf{A}\,\mathsf{D}^\star(x)$. But I have doubts as to whether we really understand this. In any case, if we do add such an operator, we must be careful

[34] 'Arbitrarily high levels' means 'for as large ordinals as there are notations'; for higher ordinals than that, talk of levels of indeterminacy makes no clear sense.

[35] Allowing this is effectively the same as allowing universal quantification over the σ, but not allowing such universal quantifications to be embeddable in other contexts like negations. But whereas a restriction on embedding seems grossly *ad hoc*, the use of schematic reasoning has a natural rationale.

that the combination of laws we postulate for it does not breed paradox, and the paradox-free ways of introducing such a $\mathsf{D}^\star$ are not altogether attractive. Option 1: Suppose we want both the inferential rules $\mathsf{D}^\star A \models A$ and $A \models \mathsf{D}^\star A$ (perhaps strengthening the first to $\models \mathsf{D}^\star A \rightarrow A$); if so, and we also keep the rule of disjunction elimination as a general logical law, then we must reject excluded middle for a sentence $Q^\star$ that asserts that it is not $\mathsf{D}^\star$-true.[36] But *if being forced to reject excluded middle shows a kind of indeterminacy*, then any determinateness operator or defectiveness predicate subject to the rules $\mathsf{D}^\star A \models A$ and $A \models \mathsf{D}^\star A$ thus embodies a kind of indeterminacy (in a background logic like K_3^+ that allows for reasoning by disjunction-elimination). And this kind of indeterminacy (if that is the right way to think of it)[37] isn't expressible by $\neg\mathsf{D}^\star$: to assert $\neg\mathsf{D}^\star\mathsf{True}(\langle Q^\star\rangle)$ would be tantamount to asserting $Q^\star$ and thus lead to paradox. Thus $\mathsf{D}^\star$ wouldn't seem to be the most powerful determinateness operator after all, in which case the introduction of the notion would be self-defeating. Option 2: That problem could be avoided by giving up the rule $A \models \mathsf{D}^\star A$; we can then keep $\models \mathsf{D}^\star A \rightarrow A$, disjunction elimination, and excluded middle for sentences beginning with $\mathsf{D}^\star$, for $\mathsf{D}^\star\mathsf{True}(x)$ becomes very much like the unadorned truth predicate in classical theories with $(T^{\star\star})$. The problem with classical theories containing $(T^{\star\star})$ was that in them we have to assert certain claims and then deny that thet are true, which seems decidely odd. Under Option 2 we don't have precisely that problem (indeed, whenever we can assert A, we can assert that it *is* true), but we have something uncomfortably close to it: we assert of specific sentences A of our own theory that they are not $\mathsf{D}^\star$-true, which raises the obvious question "If you think those specific A aren't $\mathsf{D}^\star$-true, why did you assert them?" For this reason, I think that Option 2 should be avoided.

Option 1 *may* not be beyond defense. Defending it would require refusing to accept that the "$\mathsf{D}^\star$-liar" $Q^\star$ is either defective or non-defective; indeed, it would require having degree of belief 0 in this disjunction. So we would believe the equivalent disjunction $Q^\star \vee \neg Q^\star$ to degree 0, and would not be able to explain our doing so in terms of a belief that its disjuncts were defective. (Or in any sense "quasi-defective": for the point of $\mathsf{D}^\star$ has been stipulated to be that there is no further such notion.) The rationale for this would be that to assert $Q^\star \vee \neg Q^\star$ requires more than that $Q^\star$ be non-defective, it requires that *it be assertable that* it is non-defective; so the inability to assert the non-defectiveness of $Q^\star$ is enough to explain our inability to assert $Q^\star \vee \neg Q^\star$ without commitment as to whether $Q^\star$ *actually is* non-defective. As I say, this *may* be defensible; but it is better, I think, to regard the operator $\mathsf{D}^\star$ as not fully intelligible and

[36] The assumption of $\mathsf{D}^\star Q^\star$ leads directly to paradox, as does the assumption of $\neg\,\mathsf{D}^\star Q^\star$; so by disjunction-elimination, so does the assumption of $\mathsf{D}^\star Q^\star \vee \neg\,\mathsf{D}^\star Q^\star$; but that assumption is an instance of excluded middle.

[37] It may not be, for the italicized supposition is in fact dubious. The reason: because of higher order indeterminacy, excluded middle must fail for claims of defectiveness, so the degree of belief that a claim is defective and the degree of belief that it is non-defective need not sum to 1. I've equated the degree of belief in A's non-defectiveness with the degree of belief in $A \vee \neg A$; that means that I can't in general equate the degree of belief in A's defectiveness with 1 minus the degree of belief in $A \vee \neg A$, but can only say that the degree of belief in A's defectiveness *is no greater than* 1 minus the degree of belief in $A \vee \neg A$. So though we may reject $Q^* \vee \neg Q^*$, indeed believe it to degree 0, it does not follow that the degree of belief in the defectiveness of Q^* need be high.

to simply make do with a hierarchy of determinateness operators.

8 Revenge (2)

But don't we already have a unified defectiveness predicate, viz., 'has ultimate semantic value $\frac{1}{2}$'? And a predicate corresponding to a unified determinateness operator, viz. 'has ultimate semantic value 1'? These predicates needn't even be added to the language: they are already there, at least if (i) the base language from which we started the construction in Section 2 included the language of set theory and (ii) the model M_0 for the base language from which the construction started is definable in the base language;[38] for then the set-theoretic construction of Section 2 can be turned into an explicit definition of these predicates in terms of the vocabulary of the base language. And since we have assumed excluded middle for the base language, the definitions show that excluded middle must hold for attributions of semantic value based on these definitions. Can't we then re-institute a paradox, based on sentences that attribute to themselves a semantic value of less than 1?

No we can't, and the construction of Section 2 shows that we can't: our construction yields the consistency of the claim

$$\mathsf{True}(\langle ||A|| \neq 1\rangle) \leftrightarrow ||A|| \neq 1,$$

and of all other instances of (T) in the language. (Indeed, it shows that we can add all such claims to the base theory, without disrupting any given model of that theory.) But it may seem puzzling how paradox has been avoided. Why isn't the sentence that attributes to itself a value less than 1 paradoxical?

The answer to this has more to do with the limitations of explicit definition than it does with the notions of truth and determinacy. Let L be a language that includes the language of classical set theory. L may contain terms for which excluded middle is not valid, but let us assume that the set-theoretic portion of L is classical: excluded middle (and other classical axioms and rules) all hold for it. Tarski's undefinability theorem shows that if we insist on *explicitly defining* a predicate, say 'is a sentence of L that has ultimate semantic value 1', in classical set theory, then the defined predicate can't correspond to any normal notion of truth: it gives the intuitively wrong results *even for the classical part of L*. More fully, there will be a sentence A of classical set theory, *not containing* 'True', such that we can prove that either $A \wedge \neg(||A|| = 1)$ or $(||A|| = 1) \wedge \neg A$. (With minimal additional assumptions, ones that are met by our definition of semantic value, we can indeed specify a sentence B such that one disjunct fails when A is taken to be B and the other fails when A is taken to be $\neg B$.)[39] Thus the notion of ultimate semantic value we've defined in classical set theory doesn't fully correspond to the intuitive notion of truth *even for sentences that are in no way indeterminate or paradoxical*. (The reason is that in order to give a *definition* of semantic value we have to pretend that the quantifiers of

[38] Condition (ii) will be met for any starting model that is at all natural (provided that conbdition (i) is met).

[39] The required assumptions are that every sentence of classical set theory gets one of the values 0 and 1, and that $||\neg A||$ is 1 when $||A||$ is 0 and 0 when $||A||$ is 1.

the language range only over the members of a given set, viz. the domain of the starting model, rather than over absolutely everything. What we've defined should really be called 'ultimate semantic value *relative to the particular starting model* M_0'.) That's why in constructing a theory of truth we need to add '**True**' as an undefined predicate.

Since 'having ultimate value 1 relative to M_0' doesn't quite correspond to being true, even for sentences that are in no way indeterminate or paradoxical, it also doesn't quite correspond to being determinatly true; for as applied to sentences of this sort, truth and determinate truth coincide. And that is why sentences that assert that their own ultimate value is less than 1 (relative to M_0) are not genuinely paradoxical: no one who clearly thought through the limitations of Tarski's Theorem should have expected 'having semantic value 1 relative to M_0' to precisely correspond to any intuitive notion of determinate truth. (In particular, it is clear from Tarski's Theorem that one of the rules

$$(||A||_{M_0} = 1) \models A$$

and

$$A \models (||A|\,|_{M_0} = 1)$$

must fail *even for sentences of set theory that don't contain notions like truth or determinateness*—indeed, both rules must fail—so the notions of truth and determinateness are in no way to blame. Without such rules, the argument for inconsistency collapses.)

I've argued that we can't within classical set theory define any notion of determinate truth that fully corresponds to the intuitive notion (if there is a unique intuitive notion). This is unsurprising; for even if there is a unique intuitive notion, it doesn't obey excluded middle and so couldn't possibly be defined within a classical language. In any case, any explicit definition of 'has semantic value 1' in a classical metalanguage is bound to make this notion demonstrably fail to correspond to any reasonable notion of determinateness, for the reasons just reviewed. But that raises the question: what is the point of explicitly defining semantic value within classical set theory? I addressed this question in the context of vagueness at the beginning of Section 5, but it is worth going through a parallel discussion here.

One very important function of explicitly defining 'has semantic value 1 (relative to starting model M)' within classical set theory is that doing so enables us to increase our confidence that we have the principles of the non-classical logic right. For we have an immense amount of experience with classical set theory, enough to make us very confident of its ω-consistency, and hence of the correctness of its claims of consistency. The semantics I've provided for truth theory, despite its distortions, gives a proof within classical set theory of the consistency of naive truth theory in a nonclassical logic, so we know that that logic is indeed consistent. Indeed, the particular form of the proof shows that the resulting theory is far more than just consistent, it is "consistent with any arithmetically standard starting model"; "conservative", in one sense of that phrase. That implies that it is ω-consistent; it also implies that inconsistency can't be generated by combining it with a consistent 'True'-free theory, e.g.

a consistent theory about the history of the Paris Commune. The explicit definition of semantic value (relative to a starting model) thus serves a very important logical purpose, even if (as we've seen must be the case) there is no way to turn it into an absolute definition of semantic value that completely corresponds to any intuitive notion that relates to the full universe.

I don't mean to suggest that the only function of the semantics is to give a consistency proof for naive truth theory in the logic. After all, it is certainly not the sole demand on a logic that it make naive truth theory consistent; the logic should also be intuitive, and there is no doubt that finding a notion of semantic value in terms of which the valid inferences can be characterized can play a substantial role in making the logic intuitive. If one starts the construction from a model M_0 that is sufficiently natural (e.g. it is just like the real universe except for not containing sets of inaccessible rank), then the defined notion of having semantic value 1 relative to M_0 is very close to what we'd want of an intuitive notion of determinate truth (e.g. in the parenthetical example, it gives intuitive results as applied to any sentence all of whose quantifiers are restricted to exclude sets of inaccessible rank); so inferences that preserve value 1 will be valid in a fairly natural sense. In my view, this fact does a great deal to help make the logic intuitive. It is hard to see how we could do much better in defining determinate truth for this logic in a classical metalanguage, given that the notion of determinate truth (if there is a unique such notion) is non-classical.

This does raise an interesting question: mightn't we develop a non-classical set theory or property theory without excluded middle, and within it define a semantic value relation which is appropriate to the full universe (rather than just to the domain of a classical starting model), but which is otherwise closely analogous to the definition of the model-relative notion of semantic value given earlier? And mightn't this both validate the same logic, and be such that the defined notion 'has real semantic value 1' completely corresponds to "the intuitive notion of determinate truth"? The first part of this can definitely be done: naive property theory (property theory with unrestricted comprehension) is consistent in the same extension of Kleene logic used for the semantic paradoxes: see [7]. (As I briefly discuss there, there is some difficulty in extending this to naive set theory (i.e. adding an extensionality principle), though I don't completely rule out that this can be done.) But as for the program of using such a theory (whether the non-extensional one or the contemplated extensional one) for a semantics closely modeled on the one given here,[40] I'm skeptical: the notion of having real semantic value 1 would seem to have to be a unified determinateness predicate much like the $\mathsf{D}^\star\mathsf{True}(x)$ of the previous section; and while this might not raise difficulties about validating the full logic advocated in this paper *as regards sentences not containing that predicate*, it would cause difficulties about validating the application of some of the rules to sentences that do contain that predicate.[41] (A similar reservation arises for the idea of simply

[40] As opposed, e.g., to a "homophonic semantics" based on claims like

$\mathsf{True}(\langle A \rightarrow B\rangle) \leftrightarrow [\mathsf{True}(\langle A\rangle) \rightarrow \mathsf{True}(\langle B\rangle)]$

or

$\mathsf{True}(\langle DA\rangle) \leftrightarrow D(\mathsf{True}(\langle A\rangle))$;

and also as opposed to a conceptual role semantics of some sort.

[41] An alternative to modelling the notion of having semantic value 1 on a unified determi-

taking a notion of having real semantic value 1 as an undefined primitive and using it in a semantics; the semantics might be OK for the language without the addition, but not for the expanded language.)

How worrisome is this? Not very. As I already observed in note 25 and as countless others have observed before me, a semantic theory expressed in a given logic can explain or justify that logic only in a very minimal sense: the "explanation" or "justification" it gives of its logical principles will not only employ those very logical principles, it will usually do so in a grossly circular fashion. Logic must stand on its own; the role of a formal semantics for it (insofar as it goes beyond being a "merely algebraic tool" for consistency proofs), especially a non-homophonic one (see note 40), is as a heuristic. We who use a language L governed by the logic I'm recommending get a pretty good though not perfect heuristic for that logic without expanding the language beyond L, and indeed by using merely the classical sub-part of L. If we insist upon a heuristic that avoids the problem of the classical one then we probably do need to expand L a bit; perhaps this would be done by adding a relation of "real semantic value" that is not in L but is governed by the same logic as L. If the use of the logic of L had to depend on the semantics that might be worrisome, but it doesn't. Indeed, we don't even need to regard the heuristic as completely intelligible: cf. frictionless planes.

Unlike the notions of truth and determinateness, the notion of semantic value is a technical notion of formal semantics (sometimes a technical notion for giving consistency proofs, sometimes a technical notion for heuristic "explanations" of logical principles, sometimes both). The notion of truth, on the other hand, is certainly not a mere technical notion: as is well known ([21], Ch 1; [16]), the role of a truth predicate is to serve as a device for conjoining and disjoining sufficiently regular sets of sentences not otherwise easily conjoined and disjoined, and this role (which has great importance in everyday life) has little or nothing to do with formal semantics. To serve this role, what is needed of a notion of truth is that it adhere as closely as possible to the naive theory of truth, and the truth theory in this paper adheres to that completely. Settling for only a model-relative notion of truth would be a huge defeat, in a way that settling for only a model-relative notion of semantic value is not.

The role of the notion of determinateness is also not especially technical: we would like to be able to assert the defectiveness of certain sentences in the language, such as Liar sentences and certain sentences that crucially employ vague terms, and the determinateness predicate allows us to do so. (It isn't something we need to add on to the language, we've seen that we get it for free once we have a conditional suitable for reasoning in absence of excluded middle and for expressing the naive theory of truth.) We've also seen that a single notion of determinateness can be iterated to express the defectiveness even of sentences (e.g. extended Liar sentences) that involve the notion of determinateness. Unfortunately we can't get a single unified notion of defectiveness, but must rest content with an increasing hierarchy; but I've argued that to be less

nateness predicate would be to suppose it hierarchical: i.e., suppose that though the inference from '$\langle A\rangle$ has semantic value 1' to '$\langle\langle A\rangle$ has semantic value 1$\rangle$ has semantic value 1' is valid, still the second is stronger in that the conditional with the first as antecedent and the second as consequent is not valid. But this too would cause difficulties about validating the application of some of the rules to sentences that contain the predicate 'has semantic value'.

problematic than a hierarchy of truth predicates would be. There's no need to go beyond the language L except perhaps for introducing a heuristic for the logic we employ; and we don't even need to do so for that, if we are satisfied by the heuristic given by the classical semantics.

9 Conclusion

The logic I've proposed for dealing with the paradoxes seems quite satisfactory, but there could well be others that are equally satisfactory, or better. (I think the good ones are all likely to be "similar in spirit", in that they will be obtained from the Kleene logic K_3^+ by adding a new conditional, in a way consistent with the naive theory of truth.) I'd like to believe that the good ones can all be unified with an adequate logic of vagueness, by a unified semantics something like the one suggested in Section 5 for the logic of paradoxes proposed here. For we've seen some rather general considerations that suggest the naturalness of unification:

- The semantic paradoxes seem to arise from the fact that the standard means for explaining 'True' (namely, the truth schema) fails to uniquely determine the application of the term to certain sentences (the "ungrounded" ones); and this seems to be just the sort of thing that gives rise to other sorts of vagueness and indeterminacy.

- For both the semantic paradoxes and for vagueness, it seems important to give up certain instances of the law of excluded middle. In the case of the paradoxes, this is required if we are to consistently maintain the naive theory of truth; in the case of vagueness, we must do so to resist the view that for every meaningful question there is a fact of the matter as to its answer.

- For both cases, giving up excluded middle doesn't involve *denying* it, but rather, *rejecting* it, in a sense that requires explanation. (I believe that the account of rejection offered in Section 3, in terms of degrees of belief, is adequate to both cases. In the case of the semantic paradoxes that don't turn on empirical premises, there is little need to invoke degrees of belief other than 0 and 1; which is why the discussion of degrees of belief ended up playing little role in my discussion of the semantic paradoxes.)

- In both cases we need a reasonable conditional, not definable within Kleene logic. This is needed to allow for natural reasoning. It is also needed for more particular purposes: in the case of the semantic paradoxes, it is needed if we are to maintain the standard truth schema; in the case of vagueness, it is needed to allow the representation of penumbral connections between vague concepts, and also needed for expressing such obvious truths as that if 947 is nearly 1000 then so is 953.

- In both cases we need to be able to express not only a notion of truth (obeying the naive truth schema) but also a notion of determinate truth; so we need a determinately operator D. In both cases it is fairly natural to try to define it from the conditional. And in both cases it is natural to expect arbitrarily high orders of indeterminacy, at least in the following

sense: no matter how far we iterate D—even into the transfinite, using the truth predicate—we never get excluded middle back. In other words, for any σ, we will never have $\mathsf{D}^\sigma A \vee \neg \mathsf{D}^\sigma A$ as a general law.

Do we get arbitrarily high orders of indeterminacy in the stronger sense, that for each σ there are sentences A and possible circumstances under which we can assert $\neg \mathsf{D}^{\sigma+1} A$ but cannot assert $\neg \mathsf{D}^\sigma A$? As we've seen, we get that in the case of the semantic paradoxes, so it is consistent with the generalized semantics. I doubt, though, that we want it in the case of vagueness or ordinary indeterminacy, and I doubt that any choices of the generalized models of Section 5 that are natural to employ in these cases will yield the result. It seems to me that the best way to represent higher order vagueness in a classical model is simply to make it the case that no matter how far you reiterate the determinately operator D, you never bring back excluded middle.

As I noted early in Section 5, a model for vagueness in a classical metalanguage is inevitably somewhat distorting. The most obvious distortion is that it assigns each sentence exactly one of the semantic values 1, $\frac{1}{2}$, or 0. It is thus natural to think that when a sentence A gets value $\frac{1}{2}$, then even when the semantics does not justify the assertion of $\neg DA$, then the semantics is really treating the sentence as indeterminate; in which case the semantics is drawing a sharp line between those sentences it treats determinate and those it doesn't. If so, then the semantics does, in a sense, rule out higher order vagueness, despite the fact that excluded middle can fail for sentences of form $\mathsf{D}A$ (and indeed, for sentences of form $\mathsf{D}^\sigma A$ for arbitrarily high σ).

But this problem (which is an analog for vagueness of the version of the revenge problem discussed in the previous section) seems simply to be the inevitable result of using a classical metalanguage to do the semantics. As noted both in Sections 5 and 8, the only way we could hope to get a formal semantics that portrays a nonclassical language without distortion is to give that semantics in a nonclassical metalanguage. Whether even the use of a non-classical metalanguage would enable us to develop a semantics (of a basically truth-theoretic sort) that is both more informative than a merely homophonic semantics and in no way distorting is a matter on which I take no stand. (I have expressed some skepticism as to whether such a semantics is needed: it certainly isn't needed for an understanding of the language.) In any case, it would be hard to give the semantics in this way prior to getting a grasp on what the logic to be used in the semantic theory (and in the object language) ought to be; that's why in this paper, in exploring what the logic ought to be, I have restricted myself to the use of a classical metalanguage.[42]

[42] Thanks to Steve Yablo for very useful comments at the conference, and to Graham Priest for a helpful critique of my discussion of revenge problems that led to a substantial elaboration. The line on vagueness that I take here grew out of a sequence of unsatisfactory attempts to elaborate on the classical logic account offered in chapter 10 of [5]. I'm indebted to more people than I can name for their comments and criticisms of those earlier attempts, but in addition to those mentioned in note 19, I'd like to single out Joshua Schechter for detailed and helpful comments at several stages along the way.

References

[1] Ross T. Brady. The non-triviality of dialectical set theory. In Graham Priest, Richard Routley, and Jean Norman, editors, *Paraconsistent Logic: Essays on the Inconsistent*, pages 437–470. Philosophia Verlag, 1989.

[2] Michael Dummett. The justification of deduction. In Michael Dummett, editor, *Truth and Other Enigmas*. Harvard University, 1978.

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

[4] Solomon Feferman. Reflecting on incompleteness. *Journal of Symbolic Logic*, 56:1–49, 1991.

[5] Hartry Field. *Truth and the Absence of Fact*. Oxford University Press, Oxford, 2001.

[6] Hartry Field. Saving the truth schema from paradox. *Journal of Philosophical Logic*, 31:1–27, 2002.

[7] Hartry Field. The consistency of the naive(?) theory of properties. In Godehard Link, editor, *One Hundred Years of Russell's Paradox - Logical Philosophy Then and Now*. Walter de Gruyter, 2003.

[8] Hartry Field. Is the liar sentence both true and false? In JC Beall and Brad Armour-Garb, editors, *Deflationism and Paradox*. Oxford University Press, 2003.

[9] Hartry Field. A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic*, 32, 2003.

[10] Kit Fine. Vagueness, truth and logic. *Synthese*, 30:265–300, 1975.

[11] Harvey Friedman and Michael Sheard. An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, 33:1–21, 1987.

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

[13] Petr Hajek, Jeff Paris, and John Shepherdson. The liar paradox and fuzzy logic. *The Journal of Symbolic Logic*, 65:339–346, 2000.

[14] Saul Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72:690–716, 1975.

[15] Shaughan Lavine. *Understanding the Infinite*. Harvard University Press, Cambridge, MA, 1994.

[16] Stephen Leeds. Theories of reference and truth. *Erkenntnis*, 13:111–129, 1978.

[17] Isaac Levi. On indeterminate probabilities. *Journal of Philosophy*, 71:391–418, 1974.

[18] Vann McGee. How truthlike can a predicate be? a negative result. *Journal of Philosophical Logic*, 14:399–410, 1985.

[19] Richard Montague. Syntactic treatments of modality, with corollaries on reflexion principles and finite axiomatizability. *Acta Philosophica Fennica*, 16:153–167, 1963.

[20] Graham Priest. *In Contradiction*. Martinus Nijhoff, Dordrecht, 1987.

[21] W. V. O. Quine. *Philosophy of Logic*. Prentice-Hall, Englewood Cliffs, 1970.

[22] Greg Restall. Arithmetic and truth in Łukasiewicz's infinitely valued logic. *Logique et Analyse*, 139–140:303–312, 1992.

[23] Timothy Williamson. *Vagueness*. Routledge, London, 1994.