Solving the Paradoxes, Escaping Revenge

Hartry Field
New York University

It is "the received wisdom" that any intuitively natural and consistent resolution of a class of semantic paradoxes immediately leads to other paradoxes just as bad as the first. This is often called the "revenge problem". Some proponents of the received wisdom draw the conclusion that there is no hope of *any* natural treatment that puts all the paradoxes to rest: we must either live with the existence of paradoxes that we are unable to treat, or adopt artificial and *ad hoc* means to avoid them. Others ("dialetheists") argue that we can put the paradoxes to rest, but only by licensing the acceptance of some contradictions (presumably in a paraconsistent logic that prevents the contradictions from spreading everywhere).[1]

I think the received wisdom is incorrect. In my effort to rebut it, I will focus on a certain *type* of solution to the paradoxes. This type of solution has the advantage of keeping the full Tarski truth schema

(T) $True(\langle A\rangle) \leftrightarrow A$

(and more generally, a full satisfaction schema). This has a price, namely that we must restrict both the law of excluded middle and the law connecting $A \rightarrow B$ with $\neg A \vee B$, but we can carve the restrictions narrowly enough so that ordinary reasoning (e.g. in mathematics and physics) is unaffected.[2] I'll call solutions of this type *G-solutions*. (If you want to think of the 'G' as standing for 'good' I won't stop you.) The literature contains several demonstratively consistent solutions of this sort; for purposes of this paper there is no need to choose between them. (I will give an informal introduction to this type of solution in sections 1, 3 and 4, and a formal account in section 5.) It will turn out that any such solution generates certain never-ending hierarchies of sentences that may seem "increasingly paradoxical" (roughly speaking, it is harder to find a theory that satisfactorily treats later members of the hierarchy than to find one that satisfactorily treats earlier members); but the G-solution gives a consistent treatment of each member of each such hierarchy. The existence

[1] This latter view is only reasonable if "revenge" is less of a worry for inconsistent solutions to the paradoxes than for consistent ones. I think myself that advocates of inconsistent solutions face a *prima facie* revenge problem, and doubt that they can escape it without employing the devices I suggest in this paper on behalf of certain consistent solutions. But that is a matter for another occasion.

[2] Also, the law for the conditional will hold whenever excluded middle holds for its antecedent and consequent.

of these hierarchies prevents certain kinds of revenge problems from arising: certain attempts to state revenge problems simply involve going up a level in a hierarchy all levels of which have been given a non-paradoxical treatment.

Still, there are certain "vindictive strategies" (strategies for trying to "get revenge" against solutions to the paradoxes) that G-solutions may seem to be subject to. I'll argue that the most popular such strategy is based on a misunderstanding of the significance of model-theoretic semantics. But there is a far more interesting strategy for which this is not so. As mentioned, a G-solution generates certain never-ending hierarchies of apparently paradoxical sentences which however are each successfully treated by the account. But shouldn't it be possible to "break out of the hierarchies" to get paradoxes that are not resolved by the account? Or to put it another way: If we *can't* "break out of the hierarchies" *within the language that our solution to the paradoxes treats*, isn't that simply due to an expressive limitation in that language? That, I think, is the most difficult revenge worry for G-solutions to deal with.

Some of the worries about "breaking out of the hierarchies" turn out to be intimately connected to "the paradox of the least undefinable ordinal", a paradox in the same ballpark as those of Berry and Richard. The G-solutions provide a consistent treatment of that apparent paradox. That treatment will be an important element in my argument that we are unable to "break outside the hierarchies", but that this does not reflect an expressive limitation of the language.

Part One: Introductory Discussion

1. The Paradoxes and Excluded Middle. Imagine that we speak a first order language L: it has the usual connectives and quantifiers, and it contains no ambiguous terms and no indexicals. It is to be a very rich language, powerful enough to express all our mathematics, including the richest set theory we currently know how to develop. It should be able to talk about its own expressions and their syntax; though we needn't actually make this a separate requirement, since as Gödel showed we can use arithmetical surrogates. If we like we can also assume that L can express all our current claims about the physical world too, though this will not really matter to the problem to be discussed. Finally, L should contain terms like 'true' and 'true of'. For present purposes we needn't worry about how such terms apply to sentences and formulas in languages other than our own, so we may as well assume that they have been restricted to apply only to the sentences and formulas of L.

These assumptions about our language L are enough to generate paradoxes (or rather, apparent paradoxes). Some of them, like the Liar paradox, arise from the fact that by any of a number of well-known routes we can construct self-referential sentences: sentences that attribute to themselves any property you like. For instance, the Liar paradox arises from any sentence that directly

or indirectly asserts its own untruth; let Q be some such sentence, and $\langle Q \rangle$ its standard name.[3] Since Q asserts its own untruth, it certainly seems that

$$Q \leftrightarrow \neg True(\langle Q \rangle)$$

had better be part of our overall theory. In addition, it seems that our theory of truth ought to include every "Tarski biconditional", i.e. every instance of the schema (T) mentioned earlier; hence in particular,

$$True(\langle Q \rangle) \leftrightarrow Q.$$

But then if the conditional, and the biconditional defined from it in the obvious way, are at all reasonable, we can infer

(*) $\quad True(\langle Q \rangle) \leftrightarrow \neg True(\langle Q \rangle)$.

And being of form $B \leftrightarrow \neg B$, this leads to contradiction in classical logic.

There are similar paradoxes that don't require the construction of self-referential sentences. For instance, just as our theory of truth ought to include the instances of Schema (T), so our theory of satisfaction ought to include the instances of the following schema:

(S) For all x, x satisfies $\langle A(v) \rangle$ if and only if $A(x)$.

(where to say that x satisfies $\langle A(v) \rangle$ is the same as saying that $\langle A(v) \rangle$ is true of x).[4] In the special case where $A(v)$ is the formula 'v does not satisfy v', this yields that for all x,

$\langle v$ does not satisfy $v \rangle$ satisfies x if and only if x does not satisfy x,

and hence

(**) $\langle v$ does not satisfy $v \rangle$ satisfies itself if and only if it does not satisfy itself.[5]

Again, (**) is of form $B \leftrightarrow \neg B$ and hence leads to contradiction in classical logic.

---

[3] There is a familiar distinction between *contingent* and *non-contingent* Liar sentences. If the sentence "Nothing written on the first blackboard manufactured in 2005 is true" is written on a blackboard that (perhaps unbeknownst to the writer) was the first to be manufactured in 2005, it is a contingent Liar: given the contingent facts about blackboard-manufacture it in effect asserts its own untruth. Non-contingent Liar sentences assert their own truth independent of such empirical facts. The present discussion applies to Liar sentences of both kinds.

[4] (S) should really be called (S_1): it is the schema for the satisfaction predicate 'satisfies$_1$' that applies to formulas with exactly one free variable, and there is an analogous schema (S_n) for each satisfaction predicate 'satisfies$_n$' that applies to formulas with exactly n free variables. But (S_1) can be taken as basic: in any language rich enough to code finite sequences we can artificially define the higher satisfaction predicates in terms of 1-place satisfaction, in a way that guarantees the schemas for the former if we have (S_1): e.g. to say that 'v_1 is larger than v_2' is satisfied by o_1 and o_2 in that order is in effect to say that 'u is an ordered pair whose first member is larger than its second' is satisfied by $\langle o_1, o_2 \rangle$.

We can similarly reduce truth to satisfaction: to say that 'Snow is white' is true is in effect to say that 'Snow is white and $u = u$' is satisfied by everything (or equivalently, by something). (T) then falls out of (S_1), so (S_1) can be taken as the sole basic schema. But it's more natural to talk in terms of truth than satisfaction, so I'll keep on talking about (T).

[5] A natural abbreviation for 'v satisfies itself' would be 'v is onanistic'. But for some reason 'v is homological' has caught on instead, with 'heterological' for 'non-onanistic'.

The "G-solutions" that I'll be considering accept these derivations of (*) and (**). But unlike "dialetheic" views (e.g. [15]), they do not accept contradictions (sentences of form $C \wedge \neg C$). So they must reject all arguments that would take us (for arbitrary B) from $B \leftrightarrow \neg B$ to a sentence of form $C \wedge \neg C$.

I think that the most revealing way of trying to argue from $B \leftrightarrow \neg B$ to a contradiction is as follows:

(i) Assume both $B \leftrightarrow \neg B$ and B. Then by modus ponens, $\neg B$; so $B \wedge \neg B$.

(ii) Assume both $B \leftrightarrow \neg B$ and $\neg B$. Then by modus ponens, B; so $B \wedge \neg B$.

(iii) Since $B \wedge \neg B$ follows both from the assumptions $B \leftrightarrow \neg B$ and B and from the assumptions $B \leftrightarrow \neg B$ and $\neg B$, then it follows from the assumptions $B \leftrightarrow \neg B$ and $B \vee \neg B$. (Reasoning by cases.)

(iv) But $B \vee \neg B$ is a logical truth, so $B \wedge \neg B$ follows from $B \leftrightarrow \neg B$ alone.

I now further stipulate that G-solutions accept both modus ponens and reasoning by cases (*aka* disjunction elimination). So they take the reasoning to be valid through step (iii).

What G-solutions question is the use of the law of excluded middle in step (iv). Unlike intuitionists, though, G-theorists take excluded middle to be perfectly acceptable within standard mathematics, physics, and so forth; it is only certain reasoning using truth and related concepts that are affected.[6] There is a verbal issue here about exactly how this point should be put. One way to put it is to say that excluded middle is literally *valid* in some contexts like mathematics, but invalid outside that domain. But it might be thought that the "topic neutrality" of logic implies that if excluded middle can't be accepted everywhere then it can't be taken as literally *valid* anywhere. Even so, this doesn't undermine the claim that it is *effectively* valid[7] in contexts like mathematics: if one accepts all instances of the schema $A \vee \neg A$ that don't contain 'true', then even if one doesn't claim that they are *logical* truths one can reason from them just as a classical logician reasons in mathematics and physics. So it really makes no difference in which of the two ways we talk.

Another argument from $B \leftrightarrow \neg B$ to a contradiction runs as follows: after step (i) as above, we conclude that $B \leftrightarrow \neg B$ entails $\neg B$ by a *reductio* rule (that

[6] Actually advocates of G-solutions *might* want to further restrict excluded middle, e.g. by disallowing its application to certain sentences containing vague concepts; and indeed it is not out of the question to regard certain mathematical concepts such as 'ordinal' as having a kind of "indefinite extensibility" that is akin to vagueness. Still, for purposes of this paper I assume that excluded middle applies unrestrictedly within standard mathematics.

Another plausible restriction of excluded middle is to sentences containing normative concepts like 'appropriate' or 'reasonable'; this is relevant to certain "doxastic paradoxes" involving, for instance, sentences asserting that it is not appropriate to believe them. But such paradoxes are outside the scope of this paper.

[7] 'Effectively valid' means 'in effect valid': it has nothing to do with effective procedures. Similarly I'll use 'effectively classical' to mean 'in effect classical', i.e. excluded middle holds even if not as a logical law.

if X and B together entail $\neg B$, then X alone entails $\neg B$); that result and (ii) then give the contradiction. But the most obvious argument for that reductio rule is based on the law of excluded middle (together with reasoning by cases): the argument is that if X and B together entail $\neg B$, then since X and $\neg B$ certainly entail $\neg B$ it follows that X and $B \vee \neg B$ entails $\neg B$; and since $B \vee \neg B$ is a logical truth, this means that X entails $\neg B$. So I will assume that in giving up (or restricting) excluded middle we give up (or restrict) this *reductio* rule as well.

Admittedly, this *reductio* rule is valid in intuitionist logic even in absence of excluded middle, so I can't say that we are *compelled* to give up the reductio rule if we give up excluded middle. But intuitionist logic does not evade the paradoxes, so we had best not follow its lead.[8] My point is that there is a natural response to the Liar paradox which sees this kind of *reductio* reasoning as depending on the law of excluded middle and both as needing restriction; and that is the response that G-solutions adopt.

2. Trying to Preserve Classical Logic. Weakening classical logic to deal with the paradoxes is obviously not something to be done lightly, and there are questions about how to understand the proposal, some of which I will address in the next section. But first I'd like to briefly survey the options for handling the paradoxes within classical logic; one reason for doing this is to make the non-classical approach look more attractive, and another is to facilitate a later discussion of the "hierarchies of paradoxical sentences" that arise within G-solutions.

In classical logic, the reasoning of the Liar paradox can easily be turned into a proof of the following disjunction:

> Either
>
> (i) $\langle Q \rangle$ is true, but $\neg Q$
>
> or
>
> (ii) $\langle Q \rangle$ is not true, but Q.

At this point, classical theorists have three options. (Of course, there is also the possibility of remaining agnostic between the options, but that is of no particular interest.)

The first option is to choose disjunct (i). This would seem quite unattractive: doesn't calling $\langle Q \rangle$ true while saying "nonetheless, $\neg Q$" deprive the notion of truth of significance?

[8] Intuitionists tend to motivate the *reductio* rule by way of the law $\neg(A \wedge \neg A)$ (sometimes misleadingly called the "law of non-contradiction"). But to anyone who accepts the deMorgan law $\neg(A \wedge B) \models \neg A \vee \neg B$, this version of the "law of non-contradiction" simply amounts to $\neg A \vee \neg\neg A$, a slightly restricted version of excluded middle that few who reject excluded middle would accept. (That's why dialetheists who accept excluded middle accept $\neg(A \wedge \neg A)$, making clear that it does not adequately capture the principle that we should reject contradictions.) The intuitionist argument for *reductio* thus turns on their rejection of the deMorgan law.

The second option is to choose disjunct (ii). This seems on its face almost equally unattractive: if one holds that $\langle Q \rangle$ is not true, what is one doing holding Q?

The third option is to accept the disjunction of (i) and (ii) *while ruling out as absurd the acceptance of either disjunct.* (It is because the acceptance of either disjunct is viewed as absurd that this is really a third option, distinct from agnosticism between the first two options). This third option takes the acceptance of either (i) or (ii) to be absurd, on the ground that commitment to A *requires* commitment to A being true and conversely; but it nonetheless allows commitment to $A \vee \neg A$. Now, many people think that if one accepts a disjunction of two options each of which would be absurd to accept, one has already accepted an absurdity. Indeed, that principle appears to be built into classical logic: it is the principle of reasoning by cases (or disjunction elimination), to which attention was called above. This third option is based on rejecting that principle, except in restricted form.[9] So it is probably best thought of as only a *semi*-classical option: it does accept all the validities of classical logic, but disallows natural applications of disjunction elimination and some of the other standard meta-rules.

These three options seem to be the only possibilities for keeping the validities of classical logic without accepting contradictions.[10] Admittedly, one could insist with Tarski that the predicate 'true' should be given a hidden subscript,

[9] The restricted form is that if Γ together with A entail C *by classical rules*, and Γ together with B entail C *by classical rules*, then Γ together with $A \vee B$ entail C. The third option can accept that, but cannot accept the generalization to "entailment" by the truth rules (that commitment to A requires commitment to $True(\langle A \rangle)$ and conversely). And these truth rules must have a quasi-logical status on the third option, since it was only by holding acceptance of (i) and of (ii) to be *absurd* that the view differentiated itself from agnosticism between the first two options.

[10] I know of no one who has seriously proposed taking the first option. Classical and semi-classical logicians who do technical work on the paradoxes mostly tend to prefer the third option: see [10] and [14]; also [9], where seven of the nine types of theories discussed fall under option three. The option of choice among non-specialists seems to be option two, but some specialists prefer it as well, e.g. [2] and [13]. (If the description of the latter as a classical theory seems surprising, see [6].)

What about Kripke's seminal [11]? That's more complicated since Kripke offers a model-theoretic semantics with no instructions on how to read the theory off the semantics. But if we interpret him as suggesting that though the extension of 'True' is a fixed point, the logic is classical, then his theory also falls under option two.

An alternative and I think more attractive interpretation of Kripke is to take the set of acceptable sentences to coincide with the extension of 'True': they are both the contents of the same fixed point. But if the fixed points are based on a Kleene semantics, this gives a non-classical logic, and so is not germane to the discussion in this section. (This way of interpreting Kripke has been advocated in [20]—not altogether consistently, in my view, since Soames talks in terms of truth value gaps, which seems to presuppose the classical logic interpretation. [18] clearly distinguishes the two ways of getting a theory from a Kripkean fixed point, in the distinction between the theories there called KF and KFS.)

On the non-classical reading of Kripke, his solution is similar in sprit to the G-solutions under discussion in this paper; however, the nonclassical logic one obtains from this way of reading Kripke is unsatisfactorily weak, since Kleene semantics has no serious conditional. G-solutions do much better in this regard.

or that its extension vary with context. Still, given classical logic (even in the weak sense that includes only the validities and not the meta-rules), the above three options are the only consistent ones *when the subscript and context are held fixed*.[11]

A problem with all of the classical and semi-classical solutions is that they prevent the notion of truth from fulfilling its generally accepted role. The standard story about why we need a notion of truth ([17], [12]) is that we need it to make certain kinds of generalizations. For instance, the only way to generalize over

(Snow is white) $\rightarrow \neg\neg$(Snow is white)
(Grass is green) $\rightarrow \neg\neg$(Grass is green)],

is to first restate them in terms of truth and then generalize using ordinary quantifiers:

For every sentence, if it is true, so is its double negation.

But this says what we want it to say only if we assume the intersubstitutivity of $True(\langle A \rangle)$ with A: that is, the principle

> **Intersubstitutivity:** If Y results from X by replacing some occurrences of A with $True(\langle A \rangle)$, then X and Y entail each other. [This needs to be restricted to cases where the substitution is into contexts that aren't quotational, intentional, etc.; but I'll take the language L to contain no such contexts.]

This principle entails the truth schema in classical logic, indeed in any logic in which $A \leftrightarrow A$ is a logical truth. So the three classical and semi-classical theories all reject the intersubstitutivity principle. Thus they fail to satisfy the purpose of the notion.

For instance, we want

If everything Jones said is true then ___

to be equivalent to

If A_1 and ... and A_n then ___

on the assumption that what Jones said was exactly $A_1, ..., A_n$. This requires that the $True(A_i)$ be intersubstitutable with the A_i inside the conditional, but that won't in general be so on *any* of these theories. The semi-classical theory does *better* than the fully classical ones: it allows for intersubstitutivity of $True(\langle A \rangle)$ with A in more contexts. Indeed the fully classical ones don't even allow substitutivity for unembedded occurrences: $True(\langle A \rangle)$ and A can't be mutually entailing in a classical theory that includes disjunction elimination (as we'll see in the next section). But though the semi-classical theories do *better*,

[11] I'm putting aside solutions to the Liar paradox based on unmotivated syntactic restrictions that prevent the formation of self-referential sentences. Very strong syntactic restrictions are required for this, and the solutions are of little interest since they do not generalize to the heterologicality paradox.

that isn't good enough. An advantage of G-solutions is that not only do they accept the Tarski schema, they accept the full Intersubstitivity Principle.

I conclude this section with some further remarks on the second classical option; in particular, on a version of the second classical option that invokes a hierarchy of truth predicates. This will play a role later in the paper, where I will compare it to a hierarchy of strengthenings of a single truth predicate that is involved in G-solutions.

A common theme among proponents of the second option is that the schema (T) holds for all sentences that "express propositions", where to "express a proposition" is to be either true or false, i.e. to either be true or have a true negation. (On this view, expressing a proposition is much stronger than being meaningful: it would be hard to argue that the "contingent Liar sentences" of note 3 aren't meaningful, but they are nonetheless taken not to express propositions.) So instead of (T) we have

(RT) $\quad [True(\langle A\rangle) \vee True(\langle \neg A\rangle)] \rightarrow [True(\langle A\rangle) \leftrightarrow A]$.

It is easily seen that this is equivalent to the left-to-right half of (T), i.e. to

(LR) $\quad True(\langle A\rangle) \rightarrow A$.[12]

Obviously, then, a decent theory of truth containing (RT)/(LR) needs to contain vastly more. (It's compatible with (RT)/(LR) that nothing is true; or that only sentences starting with the letter 'B' are true; etc.) The crucial question for such a theory is, how are we to fill it out without leading to paradox?

It turns out that however we try to fill it out, we are led to the conclusion that *basic principles of the truth theory itself* fail to be true. (They also fail to be false, so that they come out as "not expressing propositions".) Of course any non-contingent Liar sentence is itself an assertion of the theory that the theory asserts not to be true, but it presumably doesn't count as one of the basic principles of the theory. But what does count as a basic principle of the theory is (RT), or its equivalent (LR). And Montague showed that with very minimal extra assumptions, one can derive from (LR) a conclusion of form

$$\neg True[\langle True(\langle M\rangle) \rightarrow M\rangle],$$

i.e. that some specific instance of (LR) isn't true. Most people regard it as a serious defect of a theory that it declares central parts of itself untrue; and saying that these parts "don't express propositions" doesn't appear to help much.

This is the point at which the idea of a hierarchy of truth predicates may suggest itself. The idea is that we don't have a general truth predicate, but only a hierarchy of predicates 'true$_\alpha$', where the subscripts are notations for ordinal numbers (in a suitably large initial segment of the ordinals that has no last member). We then agree that the principles of the theory of truth$_\alpha$ aren't true$_\alpha$, but try to ameliorate this by saying that they're true$_{\alpha+1}$. Call such a view a *stratified truth theory*.

[12] In proving a given instance of schema (RT) from schema (LR) one uses two instances of the latter, one for A and one for $\neg A$.

Besides their artificiality, stratified truth theories seriously limit what we can express, in a way that undermines the point of the notion of truth. For instance, suppose we disagree with someone's overall theory of something, but haven't decided *which part* is wrong. The usual way of expressing our disagreement is to say: not all of the claims of his theory are true. Without a general truth predicate, what are we to do? The only obvious idea is to pick some large α, and say "Not all of the claims of his theory are true$_\alpha$". But this is likely to fail its purpose since we needn't know how large an α we need. (Indeed, there would be strong pressure on each of us to use very high subscripts α even in fairly ordinary circumstances, but however high we make it there is a significant risk of it not being high enough to serve our purposes. This was the lesson of the famous discussion of Nixon and Dean in [11]. Nixon and Dean wanted to say that nothing the other said about Watergate was true, and to include those assertions of the other in the scope of their own assertions; but to succeed, each needed to employ a strictly higher subscript than the other.)

Indeed, the situation is even worse than this. For suppose that we want to express disagreement with a stratified truth theorist's overall "theory of truth" (i.e. the theory he expresses with all of his 'true$_\alpha$' predicates), but that we haven't decided which part of that theory is wrong. Here the problem isn't just with knowing how high an α to pick; rather, *no* α that we pick could serve its purpose. The reason is that it's *already part of the stratified theory* that some of its claims aren't true$_\alpha$, namely, the principles about truth$_\alpha$; that's why the theorist introduced the notion of truth$_{\alpha+1}$. So we haven't succeeded in expressing our disagreement.

The problems just mentioned are really just an important special case of a problem that I've argued to infect all classical and semi-classical theories: they can't give truth its proper role as a device of generalization. Except possibly for dialetheic theories, which I will not consider here, restricting excluded middle seems to be the only way to avoid crippling limitations on our notion of truth.

3. More on Rejecting Excluded Middle. It is important to note that in classical logic you don't need anything like the full strength of the truth schema (T) (or the satisfaction schema (S)) to derive contradictions: indeed, if you allow reasoning by cases as well as the classical validities, all that is required is the two assumptions

(T-Elim) A follows from $True(\langle A\rangle)$

and

(T-Introd) $True(\langle A\rangle)$ follows from A

(or the analogous Elimination and Introduction rules for satisfaction). For using these instead of (T), we can easily recast the derivation (i)-(iv) in Section 1 (with $True(\langle Q\rangle)$ as the B) as follows:

(i*) By (T-Elim), $True(\langle Q\rangle)$ implies Q,[13] which is equivalent to $\neg True(\langle Q\rangle)$; hence $True(\langle Q\rangle)$ implies the contradiction $True(\langle Q\rangle) \wedge \neg True(\langle Q\rangle)$;

[13] That is, Q follows from $True(\langle Q\rangle)$. (One reader took my 'A implies B' to mean 'if A then

(ii*) $\neg True(Q)$ is equivalent to Q, which by (T-Introd) implies $True(Q)$; hence $\neg True(\langle Q\rangle)$ also implies the contradiction $True(\langle Q\rangle) \wedge \neg True(\langle Q\rangle)$.

(iii*) Since $True(\langle Q\rangle) \wedge \neg True(\langle Q\rangle)$ follows both from the assumption $True(\langle Q\rangle)$ and from the assumption $\neg True(\langle Q\rangle)$, then it follows from the assumption $True(\langle Q\rangle) \vee \neg True(\langle Q\rangle)$. (Reasoning by cases.)

(iv*) But $True(\langle Q\rangle) \vee \neg True(\langle Q\rangle)$ is a logical truth, so we have a derivation of the contradiction $True(\langle Q\rangle) \wedge \neg True(\langle Q\rangle)$.

(If we strengthened (T-Elim) to the assumption of the conditional $True(\langle A\rangle) \rightarrow A$, we could give a derivation that doesn't involve reasoning by cases.)

In fact, we don't even need the full strength of (T-Introd); we can make do with the weaker assumption

(T-Incoherence) A and $\neg True(\langle A\rangle)$ are jointly inconsistent.

Inconsistency proof: $True(\langle Q\rangle)$ implies Q by (T-Elim), and $\neg True(\langle Q\rangle)$ implies Q since it is equivalent to Q, so we derive Q using reasoning by cases plus excluded middle. Using the other half of the equivalence between Q and $\neg True(\langle Q\rangle)$, we get $Q \wedge \neg True(\langle Q\rangle)$, which is inconsistent by (T-Incoherence).

The fact that the paradox arises from weaker assumptions than (T) is important for two reasons. First and most obviously, it means that if we insist on keeping full classical logic we must do more than restrict (T), we must restrict the weaker assumptions as well. But the second reason it's important concerns not classical solutions, but G-solutions: it gives rise to an important moral for what G-solutions have to be like.

For even though G-solutions take truth to obey the Tarski schema (T), we'll see that they recognize other "truth-like" predicates (e.g. 'determinately true') that don't obey the analog of (T) but do obey the analogs of (T-Elim) and (T-Introd) (or at the very least, (T-Elim) and (T-Incoherence)). For each truth-like predicate, there is a Liar-like sentence that asserts that it does not instantiate this predicate. Reasoning as in (i*) and (ii*) is thus validated, and since G-solutions accept reasoning by cases without restriction, paradox can only be avoided by rejecting the application of excluded middle to these Liar-like sentences formed from truth-like predicates. (Since excluded middle is to hold within ordinary mathematics and physics, this means that no truth-like predicates can be constructed within their vocabulary.) In short, G-solutions are committed to the view that *there can be no truth-like predicate for which excluded middle can be assumed*; as we'll see, the conviction that there *must* be truth-like predicates obeying excluded middle is one primary source of revenge worries.

I close this section by trying to make clear what is involved in restricting the application of excluded middle to certain sentences, e.g. the Liar sentence,

---

B', and on this basis accused me here of illicitly extending (T-Elim) to hypothetical contexts; but that is not why I mean by 'implies'.)

when one accepts the intersubstitutivity of $True(\langle Q\rangle)$ with Q. In particular, what is the appropriate attitude to take to the claim $True(\langle Q\rangle)$? According to the sort of solution to the paradoxes I've sketched, one must reject the claim that $True(\langle Q\rangle)$ and also reject the claim that $\neg True(\langle Q\rangle)$, since these claims each imply a contradiction relative to any theory of truth that implies the Tarski biconditionals. (One can take rejection as a primitive state of mind, involving at the very least a refusal to accept; a slightly more informative account of rejection can be found in [4] (Section 3).) We must likewise reject the corresponding instance of excluded middle

Z: $\quad True(\langle Q\rangle) \vee \neg True(\langle Q\rangle)$,

for it too leads to contradiction. And because we reject Z, our refusal to either accept $True(\langle Q\rangle)$ or accept $\neg True(\langle Q\rangle)$ doesn't seem appropriately described as "agnosticism" about the truth of Q. We would be agnostic about $True(\langle Q\rangle)$ if we believed Z but were undecided which disjunct to believe; but when we reject Z the very factuality of the claim that $True(\langle Q\rangle)$ is being put into question, so our not believing $True(\langle Q\rangle)$ while also not believing $\neg True(\langle Q\rangle)$ isn't happily described as "agnosticism".

It should be immediately noted that a solution of this sort does *not* postulate a "truth value gap" in Q: it does not say that Q is neither true nor false, i.e. that neither Q nor its negation is true. It also does not say that Q is neither true nor *not true*. Saying that Q is "gappy" or "non-bivalent" in either of these senses would trivially entail that Q is not true, which (by the Tarski biconditionals and modus ponens) leads to contradiction. *Since the claim that Q is "gappy" (non-bivalent) leads to contradiction, we must reject it.*

That isn't to say that we should believe that Q is bivalent (or that it is not "gappy"; these are the same, assuming the equivalence of $\neg\neg A$ to A, as I henceforth shall). The claim that Q is bivalent or non-gappy amounts to

Z* $\quad True(\langle Q\rangle) \vee True(\langle \neg Q\rangle)$;

this in turn amounts to Z (non-truth and falsity turn out to coincide as applied to sentences in the language), and as we've seen, Z must be rejected.

If it seems odd that both the claim that Q is gappy and the claim that it is not gappy lead to contradiction, it shouldn't: from the fact that $Gappy(\langle Q\rangle)$ and $\neg Gappy(\langle Q\rangle)$ each lead to contradiction, all we can conclude is that

Z$^{@}$ $\quad Gappy(\langle Q\rangle) \vee \neg Gappy(\langle Q\rangle)$

leads to contradiction; so the proper conclusion is that this instance of excluded middle must also be rejected.[14] In other words, *the claim that Q is "gappy" has the same status as Q itself has.* In particular, just as it is misleading to declare ourselves "agnostic" about the Liar sentence, it is also misleading to

[14] Indeed whenever one rejects a given instance $A \vee \neg A$ of excluded middle, one should also reject the instance $(A \vee \neg A) \vee \neg(A \vee \neg A)$, for they are equivalent by very uncontroversial reasoning; hence one should reject $Bivalent(\langle A\rangle) \vee \neg Bivalent(\langle A\rangle)$. [Reason for the equivalence: $\neg(A \vee \neg A)$ implies $\neg A$, so $(A \vee \neg A) \vee \neg(A \vee \neg A)$ implies $(A \vee \neg A) \vee \neg A$, which implies $A \vee \neg A$. The other direction is trivial.]

declare ourselves "agnostic" about the claim that the Liar sentence is "gappy" or the claim that it is bivalent: for we don't recognize that there is a fact to be agnostic about.

I think it would be a serious problem if there were no way to assert the "defective" status of Q within the language. As we'll see, there is way; but it can't be done by saying that Q suffers a truth value gap.

4. The Berry-Richard paradox. I think that all of the semantic paradoxes turn on excluded middle, though some of them (especially some of the ones involving conditionals) do so in an indirect and unobvious fashion. I will make this precise in Section 5. There I will introduce a language that contains a "quasi-classical conditional" which obeys many of the classical laws for conditionals even in the absence of excluded middle; moreover it reduces to the material conditional when excluded middle is assumed for antecedent and consequent. I will then state a result (proved elsewhere) according to which every semantic "paradox" *that can be formulated in this language* has a solution that is compatible with the Tarski biconditionals. The solution may depend on the failure of some of the classical laws for the conditional, but that failure will always be traceable to a breakdown in excluded middle for the antecedent or consequent of one of the conditionals in question. We thus diagnose these apparent paradoxes as only apparent, they depend on illicit applications of excluded middle.

Of course, the fact that those apparent paradoxes *that can be formulated in the language* turn out not be genuinely paradoxical does not settle the revenge issue: settling that issue requires considering the possibility of expanding the language to get new paradoxes. I will have a lot to say toward undermining the idea of revenge in later sections.

First though I will consider how the paradoxes of definability fare on this sort of view. There are a number of slightly different paradoxes of definability, for instance Berry's paradox and Richard's paradox, but they all have the same underlying idea. Because of its relevance later in the paper, I will focus attention on the following variant of the Berry and Richard paradoxes.

Recall that L is a first order language adequate to its own syntax, and that contains a satisfaction predicate. From that predicate we can define 'L-definable':

> z is L-definable if and only if there is at least one formula of L (with exactly one free variable) that is satisfied by z and by nothing else.

Now, L is assumed to be built from a finite or countably infinite vocabulary, so it contains only countably many formulas; from which it follows that only countably many things are L-definable. But there are uncountably many ordinal numbers; indeed, uncountably many *countable* ordinal numbers. So there are (countable) ordinal numbers that are not L-definable. So there is a smallest ordinal number that is not L-definable, and it must be unique. But then 'v is an ordinal number that is not L-definable but for which all smaller numbers are

L-definable' is uniquely satisfied by this ordinal, so it is L-definable after all, which is a contradiction. That is the paradoxical line of argument.

Any solution to the paradoxes of satisfaction will implicitly contain a solution to this definability paradox. On *classical logic* solutions, if the language L contains the predicate 'satisfies' then certain instances of schema (S) from Section 1 are refutable; and in particular, if we define 'L-definable' from 'satisfies' as above, there will be counterinstances to even the more restricted schema

(S_{defin}) For all x, x satisfies 'v is L-definable' if and only if x is L-definable.

This gives one possible diagnosis of the error in the argument: that it lies in the inference from σ being the uniquely smallest L-undefinable ordinal to 'v is the smallest L-undefinable ordinal' being uniquely satisfied by σ.

But on the approach that I've sketched, we are committed to maintaining all instances of (S), and in particular all instances of (S_{defin}). Where then does the reasoning of the paradox go wrong?

Where the reasoning goes wrong, I think, is that it makes an implicit application of excluded middle to a formula involving 'L-definable'. Excluded middle can be assumed for certain restricted definability predicates. For instance, let L_0 be obtained from L by deleting 'satisfies' and terms defined from it (such as 'definable') or closely related to it; then excluded middle holds for formulas that contain 'L_0-definable' (as long as they don't contain problematic terms in addition). There is no even *prima facie* problem about the least ordinal that is not L_0-definable, since the description of it just given is not in L_0. Similarly for expansions of L_0 in which the application of 'satisfies' is somehow restricted in a way that guarantees excluded middle (e.g. a language L_1 in which 'x satisfies y' can occur only in the context 'x satisfies y and y is an L_0-formula', or a language L_2 in which 'x satisfies y' can occur only in the context 'x satisfies y and y is an L_1-formula'; and given a well-defined hierarchy of such expansions, each of which includes all the vocabulary of the previous, one gets a hierarchy of definability predicates, each more inclusive than the previous. But for definability in the full language L, the fact that excluded middle must be rejected for 'satisfies' suggests that it will almost certainly have to be rejected for the predicate 'L-definable' defined from it; and the paradox shows that indeed it does.

The implicit application of excluded middle to a formula involving 'L-definable' occurred in the step from

(1) There are ordinal numbers that are not L-definable

to

(2) There is a smallest ordinal number that is not L-definable.

To see that the inference from (1) to (2) depends on excluded middle, consider any specific ordinal β, and suppose that every ordinal less than β is L-definable. Given this supposition, (2) says in effect

(3) Either β is not L-definable, or there is an ordinal $\alpha{>}\beta$ such that α is

not L-definable and all its predecessors *are* L-definable.

But this entails

(4) β is not L-definable or β is L-definable;

and so if we reject (4) we must reject (2).[15] But there is no call to reject (1): there are certainly ordinals that are not L-definable, for uncountable ones can't be L-definable (and there may well be sufficiently large countable ones which are definitely not L-definable too). So the inference from (1) to (2) relies on excluded middle.

This resolution of the paradox may seem to have a high cost. For the inference from "There are ordinals α such that $F(\alpha)$" to "There is a smallest ordinal α such that $F(\alpha)$" is absolutely fundamental to ordinary set-theoretic reasoning; doesn't what I'm saying count as a huge and crippling restriction on ordinary set theory? Not at all: ordinary set theory allows sets to be defined only by "effectively classical" properties, that is, properties F for which the generalized law of excluded middle $\forall x[F(x) \vee \neg F(x)]$ holds. I'm not suggesting any restriction whatever on the ordinary laws of set theory; what I am saying, and what is independently quite obvious, is that one has to be very careful if one wants to *extend* set theory by allowing properties (or formulas) that aren't known to be effectively classical into its axiom schemas.

This point is worth elaboration. Standard set theory (ZFC) contains two axiom *schemas* (the schemas of Separation and Replacement). On a *strict* interpretation of the theory, the allowable instances of the schema are just those instances in the language of set theory; however, the "impure" set theory that most of us accept and employ is more extensive than this, it allows instances of the schemas in which physical vocabulary occurs (e.g. we take the separation schema to allow us to pass from the existence of a set of all non-sets to the existence of a set of all neutrinos). But when the law of excluded middle is not assumed to hold unrestrictedly, there is a question of just how far the extension should go. I think a suitable extension of the schema of separation to be the rule

(Extended Separation) $(\forall x \in z)(Ax \vee \neg Ax) \vdash \exists y \forall x(x \in y \leftrightarrow x \in z \wedge A(x))$ (allowing free parameters in the formula $A(x)$), *where any vocabulary at all, including 'true', can appear in* $A(x)$. Requiring excluded middle as an assumption of separation seems reasonable: otherwise, we would license sets for which membership in the set depends on whether the Liar sentence is true; given extensionality, this would lead at the very least to indeterminate identity claims between sets, and it isn't at all clear that paradox could be avoided even allowing that. But Extended Separation as formulated above avoids such oddities, while allowing such sets as the set of true sentences of number theory (and the set of true sentences of set theory that don't contain 'true'); it seems to me as much of an extension of separation to the language containing 'true' as we ought to want.

[15] Which isn't to say that we should accept the negation of (2): that would require (an existential quantification of) a negation of excluded middle, which would lead to contradiction.

It is easy to see that if the formula $F(x)$ is allowed to contain non-classical vocabulary, then Extended Separation (together with the fact that every non-empty set has a member of least rank) justifies reasoning from "There is at least one ordinal α such that $F(\alpha)$ *and such that for all ordinals* $\beta{<}\alpha$, $F(\beta) \vee \neg F(\beta)$" to "There is a smallest ordinal α such that $F(\alpha)$". And in applications of set theory in which the formula $F(x)$ is in standard mathematical or physical vocabulary, we don't need to bother stating the italicized clause since it is always satisfied. But once we allow $F(x)$ to contain notions like 'true' or 'satisfies' or notions explained in terms of them such as 'L-definable', that clause is required: forgetting it involves an illicit assumption of excluded middle, and it is on that that the Berry-Richard paradox rests.[16]

Part Two: Model Theory and Revenge

5. Conditionals and G-Logics. It's now time to give slightly more detail about the sort of logic I have in mind for dealing with the paradoxes—a *G-logic*, I'll call it. My plan is not to specify any one such logic, but to specify a class of logics any of which would deal with the paradoxes along the lines I have sketched. The logics differ only in the details of the treatment of the conditional.

The simplest way to specify the class of G-logics is to specify a type of model-theoretic semantics for members of this class—a *G-semantics*. For any specific G-logic $\pounds$, the corresponding G-semantics will give a definition of *$\pounds$-valid inference* within classical set theory; since classical set theory is accepted both by the advocates of $\pounds$ and their classical opponents, the definition of $\pounds$-validity will be intelligible to all.

I need to make a small generalization of the usual framework for model-theoretic definitions of validity: I need to allow the size of the valuation space on which the model is based to depend on the size of the domain of the model. More fully, for any cardinal number c, we fix a value space V_c with a specific subset D_c (the "designated values" of V_c); the various V_c may all be the same, but needn't be. We then stipulate that a *c-model* M consists of a non-empty domain U *of cardinality no greater than* c for the quantifiers of the language to range over, together with an assignment of an object in U to each name of the language, an operation on U to each function symbol in the language, and a "V_c-valued extension" to each predicate in the language; where a V_c-valued extension (for an n-place predicate) is a function that assigns members of V_c to n-tuples of members of U. This apparatus (in conjunction with certain operations on V_c) will be used to assign a value $|A|_M$ in V_c to each sentence A of the language

[16] What I've said here about the "least ordinal principle" is true of the least number principle in arithmetic: it too requires an excluded middle premise. Also, it's worth remarking that even positive forms of induction (ordinary or transfinite) must be treated with care when the predicate in question is not assumed to obey excluded middle: e.g., the rule form of induction

$$\forall\alpha[(\forall\beta < \alpha)(F\beta) \rightarrow F\alpha] \models \forall\alpha F\alpha$$

is generally valid, but an excluded middle premise is required for getting the conditional from this.

("the semantic value of A in M"). We then define an inference among sentences of the language to be *c-valid* (in the given logic $\pounds$) if in *every* c-model in which the premises take on designated values of V_c, the conclusion does too. And we define it to be *valid* if it is c-valid for every cardinal number c. (The definition extends in a natural way to inferences among formulas with free variables.) As remarked above, this notion of validity is definable in classical set theory, which is something that advocates of $\pounds$ and advocates of classical logic both accept.

The idea of defining validity in a model-theoretic semantics *formulated in classical set theory* may seem to make it inevitable that some sort of paradox will arise for the account. For we've seen above that a G-solution can't allow for "truth-like" predicates to which excluded middle applies. But isn't 'has a designated value' a "truth-like" predicate? And doesn't defining it in classical set theory guarantee that excluded middle must hold for it? It may seem, then, that if we are going to pursue a G-solution to the paradoxes we must explain the logic in some other way than by a model theory given in classical set-theoretic terms. Having noted this apparent problem I will put it aside until Sections 8 and 9, where I will argue that it rests upon a misunderstanding of the nature of model theory.

Turning now to the specifics of the model theory for G-logics, what should we take the value spaces V_c to be, and how should we assign values in them to sentences? If it weren't for the conditional, we could use a simple 3-valued semantics (whatever the cardinality of the model): the values might be called 0, $\frac{1}{2}$ and 1, and we would assign one of these values to each sentence; or more generally to each formula relative to an assignment s of objects in the domain of M to its free variables. (The assignments are of course all relative to a model as well as to an assignment, and indeed, the set of assignments depends on the domain of M; but for convenience I'll omit the reference to M, and often the reference to s as well, in what follows.) We'd take the value of a conjunction (relative to s) to be the minimum of the values (relative to s) of the conjuncts, the value of a disjunction the maximum, and the value of $\neg A$ to be 1 minus the value of A. The quantifiers are analogous to conjunction and disjunction. (More precisely, the value of $\forall x A$ relative to s is the minimum of the value of A relative to all the various expansions of s obtained by assigning objects to x.) We'd take 1 as the sole designated value: that is, we'd take the valid inferences to be those that in all valuations preserve the value 1. Then no sentence and its negation can both be designated; and instances of excluded middle needn't be designated, since they can have value $\frac{1}{2}$. This is called the *strong Kleene semantics* for the conditional-free language, and the logic for the conditional-free language that it generates is called *Kleene logic*. It's the semantics Kripke mostly used in [11].

This simple approach won't work if the language is to contain a conditional validating the schemas (T) and (S), for it is not hard to see that there is no 3-valued connective behaving anything like a conditional for which the asso-

ciated Tarski biconditionals all can have value 1.[17] But the approach can be generalized: we can let the spaces V_c have many more than three values (indeed, for each cardinal c we can take V_c to have more than c values), and we can take each V_c to be only partially ordered instead of linearly ordered. I will consider only the case where each V_c has a least element 0 and a greatest element 1. I will assume that there is an "up-down symmetry" operation * on V_c: an operation that reverses order and which when applied twice to any element leads back to that element. This operation will correspond to negation. I will also require that V_c be "c-complete": that is, each set of members of V_c that has cardinality no greater than c must have a greatest lower bound and a least upper bound. (In all cases of interest the models are infinite, so c is infinite; so "c-completeness" implies "2-completeness", i.e. every pair of elements has a greatest lower bound and least upper bound. If we wanted to consider the case where $c < 2$, we'd have to add a 2-completeness requirement.) Using c-completeness (and 2-completeness), we can take the value of $A \wedge B$ to be the greatest lower bound of the values of A and of B, and the value of $\forall x A$ to be the greatest lower bound of the values of A relative to all the possible assignments of objects to x; similarly for $\vee$ and $\exists$, using least upper bounds.

We also want to guarantee that as in the 3-valued semantics, $\vee$-Elimination and $\exists$-elimination hold. I will stick to the simplest case, in which 1 is the sole designated value.[18] Then the most natural way to guarantee the validity of $\vee$-Elimination is to stipulate that the least upper bound of two elements of V_c isn't 1 unless one of those elements is 1. This guarantees $\vee$-Elimination, in the restricted form that if the inferences from A to C and from B to C are both valid then so is that from $A \vee B$ to C. And we can get from this to the more general form (that if the inferences from Γ and A to C and from Γ and B to C are both valid then so is that from Γ and $A \vee B$ to C) if we assume the distributive law; for this and other reasons, I will assume distributivity. Similar remarks apply to $\exists$-elimination: for the restricted form without side formulas Γ, we assume that if S is any subset of V_c whose cardinality is no greater than c, then 1 is not the least upper bound of S unless 1 is a member of S;[19] and we get the unrestricted form that allows for side formulas if we make a weak infinite distributivity assumption.[20]

Because of the features mentioned so far, the spaces V_c can be regarded as

[17] The 3-valued conditional that "comes closest" to adequacy is the Lukasiewicz 3-valued conditional (where $|A \rightarrow B|$ is 1 if $|A| \leq |B|$, 0 if $|A| = 1$ and $|B| = 0$, and $\frac{1}{2}$ otherwise). But even here, if C is a Curry-like sentence that asserts that if it is true then so is the Liar sentence Q, then the Tarski biconditionals for Q and C can't both get value 1.

[18] A more general approach would take the set of designated values to be any prime filter not containing both v and v^* for any v, where * is the operator corresponding to negation.

[19] In typical cases $V_c - \{1\}$ will have no maximum member. (Indeed, we'll see some reason in Section 10 to impose a condition (∂c) on a satisfactory G-semantics that would entail this.) In such cases, 1 must be the least upper bound of $V_c - \{1\}$, so the condition in the text implies that V_c must have greater cardinality than c and hence greater than the cardinality of any model that has it for a value space. This is why we must allow the value space to depend on an upper bound of the cardinalities of the models that employ it.

[20] Namely $a \sqcap (\sqcup_\alpha \{b_\alpha\}) = \sqcup_\alpha \{a \sqcap b_\alpha\}$, when $\{b_\alpha\}$ has cardinality no greater than c.

a "fine-graining" of the 3-valued semantics: in any such space, the values other than 0 and 1 simply partition the value $\frac{1}{2}$ given in the 3-valued semantics. As a result, when considering inferences among conditional-free sentences it makes no difference whether you use the 3-valued semantics or one of the V_c. So the logic governing conditional-free sentences is just Kleene-logic.

The point of the fine-graining is to handle the conditional. What we need to do is add an operator $\Rightarrow$ on the spaces V_c of fine-grained values to correspond to the $\rightarrow$. The operator should obey reasonable laws, of which the foremost is

(I) $A \rightarrow B$ should have value 1 when and only when the value of A is less than or equal to that of B.

Defining $A \leftrightarrow B$ as $(A \rightarrow B) \wedge (B \rightarrow A)$, (I) implies that $A \leftrightarrow B$ has value 1 if and only if A and B have the same value. Given that all the operators are evaluated value-functionally, this then implies

($\mathbf{I}_{Cor}$) When $A \leftrightarrow B$ has value 1 and X_B results from X_A by substituting B for one or more occurrences of A then X_B should have the same value as X_A.

I'll say that two formulas A and B are *of equal strength* if $A \leftrightarrow B$ is valid, and that A *is at least as strong as* B if $A \rightarrow B$ is valid. Because of (I), the claim that A is at least as strong as B amounts to the claim that for every model M and every c at least as big as the domain of M, $|A|_M \leq |B|_M$ in V_c. So when A is at least as strong as B, the inference from A to B is valid; the converse claim fails.

Further reasonable laws for the conditional include the following:

(II) Strengthening of the consequent should strengthen the conditional and strengthening of the antecedent should weaken it;

(III) If A has value 1 and B has value 0 then $A \rightarrow B$ should have value 0;

and probably

(IV) $A \rightarrow B$ should have the same value as $\neg B \rightarrow \neg A$.

(I) (together with the assumptions we've made about negation) already implied a weak form of (IV), viz. that $A \rightarrow B$ should have the value 1 if and only if $\neg B \rightarrow \neg A$ does; the extension (IV) seems highly natural, but will play only a tangential role in what follows.

Note that (I) and (III) together imply that when the values of A and B are restricted to the set $\{0, 1\}$ (i.e., when $A \vee \neg A$ and $B \vee \neg B$ take on value 1), then the conditional $A \rightarrow B$ is to be evaluated just like the material conditional $\neg A \vee B$, which given the restriction is bound to behave classically. So any failure of a classical law for the conditional is ultimately due to a failure of excluded middle in an antecedent or consequent.

We might expand this list of reasonable laws in various ways, but for the sake both of generality and simplicity let's leave it at that;[21] then a *DMC-semantics* ("deMorgan semantics with conditional") is any semantics based on partially ordered sets V_c with operators that satisfy the laws mentioned. There are many examples of DMC-semantics in the literature: the most famous examples are the various Lukasiewicz multivalued logics. (In the Lukasiewicz logics, the same value space—e.g. the interval of real numbers—is used for every cardinal c; and the partial order $\leq$ is in fact a linear order.)

Unfortunately, most DMC-semantics will not suffice for our needs. Indeed, we've seen that no semantics consistent with the truth schema (T) can permit a "truth-like operator" that obeys excluded middle; but in most versions of DMC-semantics, including the Lukasiewicz versions, such operators can be defined using the conditional.

Our overall goal is a DMC-semantics that is consistent with the truth schema (T), or more generally the satisfaction schema (S). Given (I_{cor}), this requires that there be models in which for each sentence A, $True(\langle A\rangle)$ has the same value as A; and for each formula A with one free variable, $Satisfies(x, \langle A\rangle)$ always has the same value as the result A_x of replacing all free occurrences of the free variable in A by 'x'.[22] That's what *consistency* with (T) and (S) require. Actually what we want is more than mere consistency, we want a kind of "conservativeness" result involving "consistency with any standard starting model". Basically what this requires is that any standard "starting model" (classical model M_0 for the fragment L_0 not containing 'True' and 'Satisfies') can be converted to a model of the DMC semantics that meets these conditions on 'True' and 'Satisfies' and whose part not involving 'True' and 'Satisfies' "looks just like" M_0.[23] (Whenever in the rest of the paper I talk of consistently adding the truth schema to a semantics, what I really mean is this.) Let a *G-semantics* be any DMC-semantics that meets this conservativeness requirement.

There are in the literature several different G-semantics: several solutions to the paradoxes of the general sort I've sketched that give all the biconditionals of

[21] For later reference I mention a strengthening of (III):

(III_s) If A has greater value than B then $A \rightarrow B$ should have value 0.

If the values were linearly ordered, this with (I) would yield excluded middle for conditional claims, which would inevitably breed paradox; but I know of no system adequate to the paradoxes whose values are linearly ordered, and the published G-solutions do all satisfy (III_s).

[22] With suitable change of bound variables in A, if x occurs in A so as to create a conflixct with the substitution.

[23] By a standard model I mean one whose syntactic part is standard (i.e. for each e_2 in the domain there are only finitely many e_1 in the domain for which $\langle e_1, e_2\rangle$ satisfies 'v_1 and v_2 are expressions and v_1 is part of v_2'); equivalently, whose arithmetic is standard. The "looks just like M_0" condition means that the two models have the same domain and the same assignments to individual constants and function symbols, and that M assigns to any n-place predicate of L_0 a function that maps any n-tuple into 1 if it is in the M_0-extension of the predicate and into 0 otherwise.

form (T) and (S) value 1.[24] Given (I_{Cor}), this means that *they also validate the intersubstitutivity of* $True(\langle A\rangle)$ *with* A*, even within the scope of other operators such as the conditional*: if X results from Y by substituting $True(\langle A\rangle)$ for one or more occurrences of A, then X and Y get the same value; so $X \rightarrow Y$ and its converse get value 1. (More generally, they validate the intersubstitutivity of $Satisfies(x, \langle A\rangle)$ with A_x even within the scope of other operators. In what follows I will frequently state my claims just for truth, leaving the generalization to satisfaction tacit.) Logics that can be based on such semantics (*G-logics*) are the ones that will be of interest in what follows.

For purposes of this paper it will be convenient to use the term 'valid' for G-logics in a very broad sense, one which counts every arithmetic truth, a large amount of set theory, and the basic principles of truth and satisfaction as "valid". To be explicit, let a *quasi-correct model of the 'true'-free fragment* L_0 of L be a model M_0 of L_0 such that for some inaccessible cardinal κ, if Ur is the subset of the domain that does not satisfy 'Set', then the set-theoretic portion of M_0 consists of the set of all (not necessarily pure) sets of rank less than κ built from urelements in Ur (together with the usual $\in$ relation on this domain). Let a *standard set-theoretic model for* L (in a given G-semantics) be a model of L with valuation space V_c for which the model for the 'true'-free fragment is a quasi-correct model of cardinality no greater than c, and in which all instances of the schemas (T) and (S) get value 1 (when syntax is developed in ZFC in a standard way). Then I will take an inference to be *valid* in the G-logic if it preserves the designated value 1 in any standard set-theoretic model of the G-logic. For future reference, I note that the following three set-theoretic principles come out valid in this sense even when extended to the full language containing 'True' and 'Satisfies': (i) the Extended Separation Schema of Section 4; (ii) the rule form of transfinite induction, mentioned in note 16, and (iii) the "choice principle" $(\forall x \in X)(\exists y)F(x, y) \vDash (\exists f)[dom(f) = X \wedge (\forall x \in X)F(x, fx)]$.[25]

6. Semantic Values and Truth. How do the semantic values in a G-semantics relate to the notion of truth? Putting aside a complication to be discussed in the next section, we can say that the following holds for "reasonable" models:[26]

- Sentences with value 1 are true;

[24] One is almost explicit in [3] and fully explicit in [5]; more mathematical details about it can be found in [21]. Another can be found in [7]; this one was inspired by [22], which is in the general spirit of a G-semantics but fails to satisfy intersubstitutivity of truth and satisfaction within conditionals. One can also modify Lukasiewicz continuum-valued semantics to get a G-solution, using the basic ideas from [3]. I suspect that there are many other possibilities as yet undiscovered. A much earlier paper offering something close to a G-semantics is [1].

[25] (i) and (ii) are completely evident given the quasi-correctness of the underlying model. (iii): if the premise to have value 1 in the model, then using the axiom of choice it must be that for some function g with domain $\{o \mid ||x \in X||_o = 1\}$, $||F(x, y)||_{o,g(o)} = 1$ for all $o \in dom(g)$. By the quasi-correctness of the underlying model, this g must be in the model, from which it follows that the conclusion must get value 1.

[26] We'll see there that *even in classical semantics*, the conditions may fail for some sentences whose quantifiers range over sets of arbitrarily high rank. You can take the discussion in this section to apply only to sentences with suitably bounded quantifiers.

- Sentences with value 0 are false (i.e., have true negations).

These claims are natural given the semantics. If a sentence has value 1 then anyone who knows that it has this status can assert it; and since the claim that A is true is equivalent to A itself, he or she can then assert that it is true. Similarly, anyone who knows the status of a sentence with value 0 can assert that it is false: for its negation must have value 1, so we can assert the truth of $\neg A$ and hence the falsity of A. This is intended only as an informal argument; that is the best that can be expected for connecting notions of these two different kinds.

How about sentences that (we know to) have intermediate values? It is sometimes assumed that they are neither true nor false, but that does not fit with what I've already stipulated, in particular with the intersubstitutivity of $True(\langle A\rangle)$ with A. We've seen this already with Liar sentences: they must have an intermediate value, but we can't assert that they aren't true since that would lead to contradiction, so we certainly can't assert that they aren't either true or false. But the point holds more generally. Since falsehood is equivalent to truth of the negation, to say of a sentence A that it is neither true nor false would be to say

$$\neg True(\langle A\rangle) \wedge \neg True(\langle \neg A\rangle).$$

But by the intersubstitutivity of $True(\langle A\rangle)$ with A, that would be equivalent to saying

$$\neg A \wedge \neg\neg A,$$

i.e. to accepting a contradiction; which is illegitimate in this logic since contradictions can never get the only designated value, viz. 1. So *even for sentences for which we can easily show that they can't have value 0 or 1*, we must reject the claim that they are neither true nor false. (Of course if we know them to have intermediate truth value, we also won't assert that they are either true or false; nor will we assert that they are either true, false, or neither true nor false.)

Is there anything we *can* say about the truth and falsity of sentences with intermediate semantic values? Yes. Some examples:

- When the value of A is less than or equal to that of B, then if A is true B is true and if B is false A is false.
- If A and B are both true, so is $A \wedge B$, and if $A \vee B$ is true then at least one of A and B is true.[27]

The claims in the second bullet each result directly from three applications of the equivalence between $True(\langle C\rangle)$ and C (together with the fact that '$\wedge$' and '$\vee$' are transcriptions of 'and' and 'or'). And those in the first bullet can be

[27] It is tempting to summarize these two bulleted laws by saying that truth is a "fuzzy prime filter" on the space of values—"fuzzy" because it is indeterminate whether any given sentence with intermediate truth value is in it.

reduced to the two that were informally argued for two paragraphs back. For if the value of A is less than or equal to that of B, then the value of $A \rightarrow B$ is 1; so by the claim of two paragraphs back, $A \rightarrow B$ is true, and hence (by the intersubstitutivity properties of truth and the fact that '$\rightarrow$' means 'if ...then') if A is true then so is B. Similarly for the falsehood claims, using the law (IV) for the conditional. (This is the only place in the paper I will rely on (IV).)

In sum, there is a lot we can say about the connection between the semantic values and truth and falsity; but we can't say of any claim with intermediate value either that it is true or that it isn't. Again, it wouldn't be appropriate to say that we're ignorant about whether these sentences are true. For we are ignorant about whether A is true when either A is true or A is not true but we don't know which; but in this case the assumption that either A is true or A is not true cannot be made.

7. Revenge Problems: Introductory Remarks. Since one of the requirements of a G-logic was that it be consistent with the instances of (T) and (S) all having value 1 (or rather, the stronger "conservativeness" requirement sketched at the end of section 5), there is no danger that there are any genuine paradoxes statable in the language: any apparent paradox statable in the language has a solution that is consistent with the instances of these schemas. Thus a vast array of apparent paradoxes (e.g. a Curry-like paradox involving a sentence C that asserts that if it's true then so is the Liar sentence) are automatically solved. But it might be argued that the language is expressively weak, in that certain notions that we can't easily do without are inexpressible in it; and that including those notions within the language would inevitably breed new paradoxes. This is the general idea behind revenge problems.

One notion to worry about is determinacy. We've seen that we can't without contradiction declare that the Liar sentence isn't true; but we nonetheless reject the Liar sentence (since it leads to contradiction), and it seems that there is some important sense in which we believe that it is not *determinately* true. But if we recognize such a sense of determinacy, then a full solution to the paradoxes must consider sentences that can include this notion as well as 'True' and '$\rightarrow$' (and all the other notions in the original language).

I agree with this, and will discuss how a G-logic can accommodate it in Section 10. Before that, though, I want to consider one particular form of the worry that I do *not* agree with. This worry involves a particular kind of notion of determinacy, one that is thought to be somehow read off the model-theoretic semantics. In Section 9 I will give two "simple revenge arguments" based on this form of the worry, and argue that they are mistaken. But first I will prepare the way for my reply, by trying to remove what I think are common misconceptions about model-theoretic semantics.

Not all revenge arguments are based on misconceptions about model theory: I will discuss what I see as a much more interesting revenge argument, not based on such a misconception, in Sections 13-20. I think there is some tendency in

discussion of these matters for the "simple" and "sophisticated" arguments to become intertwined, so it is important to deal with the simple ones before dealing with the sophisticated ones.

8. Model-theoretic Semantics. In Section 5 I sketched a model-theoretic semantics for the language L, in classical set theory. What is the value of giving such a semantics? The obvious answer, and the one I gave, is that such a semantics enables us to give a set-theoretic definition of a notion of logical validity for the language. When, as here, the language of set theory is part of L and is assumed to effectively obey classical logic (by virtue of $A \vee \neg A$ always being assumed, when A is in the language of set theory), we are using what is in effect a classical part of L to define validity for the full L.

That it is possible to adequately develop the theory of validity for L within a classical portion of L rests on a presupposition. It presupposes that excluded middle holds of logical notions like implication: that is, it presupposes that if Γ is a set of sentences and B is a sentence, either Γ implies B or it doesn't. This presupposition seems reasonable to me (though it is not beyond question); but even if it is rejected, the set-theoretic definition of validity gives a useful first approximation, that can be grasped by the advocate of classical logic as a first step toward understanding how to reason in the non-classical logic.

What if we shift from explaining validity to explaining truth? In my view, model theory plays at best a very indirect role in explaining truth. Rather, truth is directly explained by means of Schema (T), and model theory enters in only in helping us understand the logical connectives that occur in instances of Schema (T). More fully, (i) model theory gives an account of validity for the non-classical logic, which tells us a lot about how to reason with the connectives in the logic; (ii) once we come to understand how to reason in the logic we will fully understand its connectives; (iii) when we understand the connectives, together with the primitive non-logical symbols, then we will understand the sentences of the language; and (iv) that means that we will understand what it is for a sentence of the language to be true, given that we accept all instances of the schema (T).[28] So it is only through its role in explaining validity that model-theoretic semantics helps convey an understanding of truth for L-sentences.

I won't argue here for this positive view of how we understand the notion of truth for L-sentences. What I will do, though, is try to undermine the idea that a model theoretic semantics could have any more direct role to play in our understanding of truth.

The first point to be made here depends on the fact that the model theory is a model theory for a non-classical logic, but is being given within an effectively

[28] Of course, we want to be able to prove generalizations about truth that don't follow from the instances of (T) (though maybe they follow from the schema understood in a broader sense—see [8]). It is doubtful, though, that these are required for understanding the notion. Even if they are, my basic point is that our understanding of 'true' is given by the acquisition of a theory that contains it; the assumption that this theory consists only of Schema (T) is completely inessential to this basic point.

classical part of the language, namely set theory. The point is an obvious one: if we are to use a logic without excluded middle to handle the paradoxes, such instances of excluded middle as

$$True(\langle Q \rangle) \vee \neg True(\langle Q \rangle)$$

must be unacceptable (where Q is the Liar sentence). But if 'True' were defined in set-theoretic terms, we would have to accept it, given that excluded middle holds within set theory. So a model-theoretic semantics for a non-classical language can't possibly explain the notion of truth. (It also can't explain the notion of determinate truth or any other such notion; for according to G-solutions no such notion can be subject to excluded middle, as they would have to if defined in classical set theory. I will have much more to say about determinate truth in later sections.)

There is a second point with the same conclusion, and this one arises even for the model theory of classical languages. It is based on Tarski's theorem about the undefinability of 'true-in-L' in the 'true'-free portion of L. Tarski stated the theorem for classical languages only, but obviously it extends to non-classical languages if we assume that their 'true'-free portion is classical, since a definition of 'true-in-L' in the subpart L_0 of L that doesn't contain 'true' would yield a definition of 'true-in-L_0' in L_0.

How is it that we can use classical set theory to define 'the semantic value of A relative to a model', but can't use it to define 'true'? The answer is that semantic value is relative to a model, and that truth (in the intended sense, the sense that obeys the Tarski schema (T)) is not. And the crucial point about the relativity to a model is that *in a model, the quantifiers are restricted to the members of a set; they do not range over absolutely everything.* Without this, the explicit definition of semantic value would not be possible.

If we want to think of the model-theoretic semantics as telling us not just about validity but about truth, then we will have a special interest in what we might call *homophonic models*: models which assign to a name its real bearer and analogously for function symbols, and which in the classical case assign to a predicate those objects in its real extension *that are also in the domain of the model.* But truth-in-a-homophonic-model must be distinguished from truth (even in the classical context where claims about truth obey the law of excluded middle). To say that a sentence is true in a homophonic model M is to say in effect that it would be true if its quantifiers were restricted to the domain of M. That can be defined (if the model M is definable); but by it's very model-relativity it diverges from the notion of truth.

Consider a classical model for the 'true'-free part L_0 of the language L. L_0 includes standard set theory. Suppose we take a highly natural model for L_0, say the homophonic model whose domain consists of all non-sets together with all sets of rank less than the first inaccessible cardinal; call this homophonic model M_1.[29] This assumes of course that there are inaccessible cardinals; otherwise

[29] M_1 is thus "quasi-correct" (as defined at the end of Section 5) as well as homophonic.

there would be no such model, so we'd have to use a different example. But now consider the sentence 'There are inaccessible cardinals': it's true, but false in M_1, i.e. has semantic value 0 in M_1; its negation is false, but has value 1 in M_1. Having semantic value 1 in M_1 doesn't correspond to truth, or to determinate truth, or anything like that, even in the classical sublanguage L_0 of L. The point made here for M_1 applies to any other model that can be defined within set theory, by Tarski's Theorem, and this includes all models of set theory that are at all "natural". (Indeed, the point has an extension to "unnatural" models of set theory that are not set-theoretically definable: see [3], n. 24.)

The term 'model' is sometimes employed in a broader sense than I have been taking it, a sense in which we give a model by specifying a domain that needn't be a set. If this is done in the context of a theory in which we quantify over proper classes as well as sets, it obviously changes nothing important: the sentence 'There are proper classes' will come out having value 0 in the model even though it is (according to the theory) true (and even though the model is homophonic and quasi-correct). If it is done in the context of a theory in which we don't quantify over proper classes but regard 'proper class' talk as a dispensable manner of speaking to be construed in terms of language, then Tarski's undefinability theorem applies: the notion of truth or semantic value *in a proper class model* is not explicitly definable in the set-theoretic language; in reasoning with it, we are going beyond standard set theory, we are reasoning in a set theory expanded by adding a notion of set-theoretic truth. (And of course we have then left the notion of truth for arbitrary sentences in this expanded language undefined.)

One might think I am making too much of the fact that 'true-in-L_0' isn't explicitly definable in L_0, when L_0 is a classical language: after all, it is inductively definable in L_0,[30] and the problem is only that this inductive definition can't be made explicit because the quantifiers range over everything. I'm not sure why the possibility of inductive definition in the classical case should be thought to undermine what I've said, but there's no need to go into that: for it is completely irrelevant to the case of actual interest in this paper, the semantics of the non-classical languages used for the paradoxes. For in every case of which I am aware, the model-theoretic semantics used for those languages requires the quantifiers to be restricted *even in giving the inductive definition* of semantic value. This is certainly so for any model theoretic semantics that builds on a Kripke-like model theory (see [11]), for that requires an inductive construction *whose first step is an explicit definition of truth for the 'true'-free sublanguage*; if the quantifiers in the language weren't restricted to the members of a given set, the inductive specification couldn't get off the ground.

Indeed, for most G-solutions the point is even more striking. They build on Kripke's theory of truth, and thus are subject to the previous observation.

(These requirements are somewhat independent: e.g., quasi-correctness requires homophony in the set-theoretic vocabulary but not elsewhere.)

[30] Or to be pedantic, the term 'satisfies in L_0', from which 'true in L_0' is explicitly definable, can be inductively defined in L_0.

But in addition, the cardinality of the evaluation space V must (in typical cases anyway—see note 19) be larger than the cardinality of the starting model. This fact seems to me to seriously undermine the idea that we can somehow extrapolate an understanding of a model-independent notion like truth or determinate truth from the model-dependent notions: for if we were to try to somehow extrapolate from the case of models on domains with a cardinality c and valuation spaces with a cardinality $f(c)$ strictly bigger than c to the case of "models" whose domain is absolutely everything, this would seem to require a valuation space "strictly bigger than absolute infinity", i.e. not only with more members than any set but with more members than "the totality of absolutely everything". I don't think such an extrapolation possible: model-theoretic notions are one thing, truth and determinate truth are something else again.

To summarize this section, it is very dangerous to draw conclusions about truth and related notions from model-theoretic semantics, for at least two reasons: (a) because the model-theoretic semantics is in a classical metalanguage, so that excluded middle is assumed throughout; (b) because even the homophonic models falsify how the language works by taking the quantifiers to range over a certain set M (the domain of the model) rather than ranging over absolutely everything. Because of these two facts, the notion of semantic value is inevitably a somewhat artificial construction that can only be understood as model-relative, and conclusions about how sentences are to be evaluated with respect to properties that are *not* model-relative (for instance, truth, determinate truth, and so forth) are highly problematic. If such conclusions can be drawn at all, it is only with extreme caution.

9. The Simplest Revenge Arguments. I think that the points in the preceding section can be used to undermine the following two revenge arguments. (More difficult revenge arguments will be considered later on.) These two arguments don't depend much on the details of the G-semantics.

> **Simple inferential revenge argument:** According to the semantics, the space of semantic values is partitioned into two classes, the designated and the undesignated; and *the semantics assumes designatedness to be a classical notion*, that is, each sentence is either designated or undesignated. But suppose we had *in the language* the predicate 'has a designated value'. The predicate would not only obey excluded middle, it would also need to be "truth-like" in the sense of Section 1. That is, the following inferences would need to be valid:
>
> [*Des*-Elim] $\quad Designated(\langle A\rangle) \vDash A$
>
> and
>
> [*Des*-Incoher] $\quad A, \neg Designated(\langle A\rangle) \vDash \bot$ (where $\bot$ is an absurdity).
>
> Indeed, we'd probably want to strengthen the former to

$$\vDash \mathit{Designated}(\langle A \rangle) \rightarrow A$$

and the latter to

$$A \vDash \mathit{Designated}(\langle A \rangle).$$

But even without these strengthenings, [*Des*-Elim] and [*Des*-Incoher] lead to absurdity, using a "super-Liar" sentence Q_* that asserts that it doesn't have a designated value, by the reasoning near the end of Section 1. (To review: $\mathit{Designated}(\langle Q_* \rangle)$ and $\neg \mathit{Designated}(\langle Q_* \rangle)$ each imply Q_* (the first by [*Des*-Elim], the second by the definition of Q_*), so using reasoning by cases plus excluded middle we get Q_*; but then by the definition of Q_* we also get $\neg \mathit{Designated}(\langle Q_* \rangle)$, and these two claims together are absurd by (*Des*-Incoher).)

It seems that if we allow the notion of designatedness into the language, assumptions about it that appear almost inevitable lead to contradiction.

Before evaluating this argument, let's consider a semantic variant:

Simple semantic revenge argument: As in the inferential version, we argue that if we had in the language the predicate 'has a designated value' then we could form a "super-Liar" sentence Q_*; so

(*) $Q_* \leftrightarrow \neg \mathit{Designated}(\langle Q_* \rangle)$

must have a designated value. But this requires

(**) Q_* has designated value if and only if $\neg \mathit{Designated}(\langle Q_* \rangle)$ has designated value.

But if Q_* has a designated value, then $\mathit{Designated}(\langle Q_* \rangle)$ should too, and so $\neg \mathit{Designated}(\langle Q_* \rangle)$ should not have designated value. So given (**), the assumption that Q_* has designated value is absurd. Similarly, if Q_* does not have designated value, $\neg \mathit{Designated}(\langle Q_* \rangle)$ should have designated value; so given (**), the assumption that Q_* does not have designated value is also absurd. But Q_* either has a designated value or doesn't (the semantics being classical), so we are landed in an absurdity either way.

Again, it looks like something really unattractive is required, if we allow the notion of designatedness into the language.

What the proponent of either form of the argument holds, then, is that if we allow a designatedness predicate into the language that we are semantically evaluating, we have a paradox: we are led into contradiction unless we abandon an assumption that is central to the intuitive meaning of the notion. And if we don't allow such a predicate into the language that we are semantically evaluating, then our solution to the paradoxes works only because the language being evaluated is expressively incomplete.

A *possible* reply to both arguments, though not one I find at all attractive, is in terms of "levels of language". According to this possible reply, we have to introduce a whole hierarchy of value spaces $V_{(1)}$, $V_{(2)}$,... and a corresponding hierarchy of designatedness predicates; the paradoxes arise, on this view, by assuming that sentences containing a given designatedness predicate Des_α are themselves to be evaluated in terms of Des_α, when in reality they are to be evaluated in terms of $Des_{\alpha+1}$. (In the semantic version, the claim would presumably be that sentences containing Des_α are simply unevaluable in $V_{(\alpha)}$, but only in $V_{(\alpha+1)}$ and higher. In the inferential version, the claim would presumably be that the inference rule $A \vDash Des_\alpha(\langle A \rangle)$ (or $A, \neg Des_\alpha(\langle A \rangle) \vDash \bot$) must be restricted to the case where A has no Des predicate subscripted α or higher.) These solutions are formally adequate, but in invoking such a hierarchy of value spaces and of unrelated Des_α predicates they seem completely outside the spirit of G-solutions to the paradoxes.

Fortunately, such a "level of languages" approach is completely unnecessary. The proper reply to the simple arguments, I think, is that for the reasons discussed in the previous section, any intelligible 1-placed predicate of having designated value is model-relative. Let's stick to a predicate that relativizes to a *specific* model: 'is designated in the valuation based on M_0'. (It's easy to see that predicates like 'is designated in *all* valuations based on models of such and such sort' and 'is designated in *some* valuation based on a model of such and such sort' could only make things worse.) Now for these relativized predicates, there is no paradox: they are already in the language L, they obey excluded middle, they fail some of the assumptions used in the two arguments, *but that failure is in no way surprising or paradoxical precisely because of the relativized nature of the predicate*. For example, suppose the model M_0 is the homophonic model whose domain consists of those objects of rank less than the first inaccessible cardinal. Then 'There are inaccessible cardinals' is undesignated relative to that model, even though it is (determinately) true; and its negation is designated relative to that model, but (determinately) false. In short, if 'designated' is interpreted as 'designated relative to M_0' then lots of perfectly ordinary sentences (sentences of set theory, *not containing 'true' or other suspect terms*) have precisely the "paradoxical" features of the sentence that asserts its own lack of "M_0-designatedness". There simply is no paradox here.

Obviously the proponent of the simple revenge problem doesn't intend 'designated' to be understood as model-relative. The question then arises, how is it to be understood. I do not deny that it is possible to introduce into the language an operator (which I prefer to call 'determinately') with many of the features that the proponent of revenge wants, and which is *not* model-relative. Indeed, I think that such predicates are already definable in L! I will discuss this in the next section. But such predicates only breed paradox if they satisfy all the assumptions used in the derivations above. It turns out that one can get predicates that satisfy *most* of the assumptions used in the derivations above; the one place they fail is that excluded middle cannot be assumed for them. So

there is a revenge problem (of the sort considered in this section) only if there is reason to think that we can understand a notion of "designatedness" that obeys those other assumptions *plus excluded middle*.

And why assume that? I think what underlies the simple revenge problem is the thought that the model-relative designatedness predicates all obey excluded middle, so there must be an absolute designatedness predicate that does too. But this assumption seems to me completely unwarranted: one just can't assume that one can extrapolate in this way from the case of model-relative predicates, which make sense only by virtue of "misinterpreting" the quantifiers as having restricted range, to the unrelativized case where no such "misinterpretation" is in force. In the case of G-solutions, even the choice of value-space depends on the initial restriction of domain; if one tries to idealize away the restriction of domain, one is left without a choice of value space. How one is supposed to be left with an intuitive understanding of an absolute notion of designatedness (even one that can only be formulated in a richer language) is beyond my comprehension.

Part Three: Determinacy and Iterated Determinacy

10. Determinacy and "Strengthened and Weakened Liar Sentences". Once we assume a G-semantics, there is no danger that there are any genuine paradoxes statable in the language L: any apparent paradox statable in the language has a solution that is consistent with all instances of (T) and (S). If there is a revenge worry, it is that the language is expressively weak, and that there are concepts we need that if added to the language would breed new paradoxes. For instance, the Liar sentence is clearly somehow "defective", but we've seen that we can't explain its defectiveness as its being neither true nor false; can we explain this in some other way? It's natural to say that its defectiveness consists of its being neither *determinately* true nor *determinately* false. We can take 'determinately' to be an operator D taking formulas to formulas; from that we can form predicates of determinate truth and determinate falsity, viz., $D[True(x)]$ and $D[True(neg(x))]$ (and analogously, of determinate satisfaction).

A worry is that if we add such a determinacy operator to the language (in a way that allows us to say of the Liar sentence that it is neither determinately true nor determinately false), we will inevitably be led to new paradoxes that cannot be consistently treated. A weaker worry is that the motivations for introducing a notion of determinacy will eventually lead to a hierarchy of richer and richer languages. I will argue that both worries are unfounded.

We want to allow the operator D to apply to formulas that themselves contain D. (This is a precondition to avoiding a hierarchy of languages.) The key to avoiding paradox is that excluded middle won't be assumed for claims of determinate truth: we *don't* want that every sentence is either determinately true or not determinately true. (Analogously for determinate satisfaction.) Because of this and the fact that the semantics is given in classical logic, we can't

straightforwardly define the notion of determinate truth in terms of the semantics.

Instead, let's impose some conditions that a reasonable determinacy operator should satisfy. From a model theoretic viewpoint it seems quite reasonable to assume that D is value-functional and satisfies the following:

(a) If $|A|_{M,s} \leq |B|_{M,s}$ then $|DA|_{M,s} \leq |DB|_{M,s}$.

(b) If $|A|_{M,s} = 1$ then $|DA|_{M,s} = 1$

($\mathbf{c}_w$) $|DA|_{M,s} \leq |A|_{M,s}$

and probably the strengthened form of that

(c) If $0<|A|_{M,s}<1$ then $|DA|_{M,s}<|A|_{M,s}$ and if $|A|_{M,s} = 0$ then $|DA|_{M,s} = 0$

(a), (b) and (c_w) correspond to natural inferential principles: (a) to the principle $A \rightarrow B \vDash DA \rightarrow DB$, (b) to $A \vDash DA$, and (c_w) to $\vDash DA \rightarrow A$. The inferential content of the remainder of (c) is that $A \rightarrow DA \vDash A \vee \neg A$. (The converse inference $A \vee \neg A \vDash A \rightarrow DA$ follows from (b).)[31]

Conditions (a), (b) and (c_w) are clearly insufficient for counting as a determinacy operator, for they are compatible with D being the identity operator. (c) partially rectifies that, but is insufficient to guarantee that we can legitimately declare the Liar sentence Q to be not determinately true (or indeed, to guarantee that we can declare $\neg DA$ for any A for which we cannot declare $\neg A$). Model theoretically, what we need is that DQ have value 0. For that, the following is sufficient (and in most versions of G-semantics, also necessary):

(d) If $|A|_{M,s} \leq |\neg A|_{M,s}$ then $|DA|_{M,s} = 0$,

which corresponds to the "modified *reductio* principle" $A \rightarrow \neg A \vDash \neg DA$.[32]

If one takes the model theory sufficiently seriously one may want to replace (a)-(d) by a slightly stronger condition:

(e) There is an operator ∂ on the space V_c of values such that if $|A|_{M,s}$ is v, $|DA|_{M,s}$ is ∂v, and which satisfies analogues of (a)-(d). That is, which satisfies

(∂a) If $v_1 \leq v_2$ then $\partial v_1 \leq \partial v_2$

[31] It doesn't follow that $A \rightarrow DA$ is fully equivalent to excluded middle, i.e. that $|A \rightarrow DA| = |A \vee \neg A|$, and in typical G-logics (e.g. those meeting condition (III_s) of note 21) this can fail: for instance, where Q is the Liar sentence, $Q \rightarrow DQ$ may have value 0, which can never be the case for an instance of excluded middle.

[32] **Sufficiency:** Since $|Q| = |\neg True(\langle Q \rangle)| = |\neg Q|$, (d) requires that $|DQ| = 0$. **Necessity:** If $|A|_s \leq |\neg A|_s$ then $|A|_s \leq |A \wedge \neg A|_s$. In every G-semantics I know of (e.g. that of [5]), we have that for any A, $|A \wedge \neg A|_s \leq |Q|$. Assuming that, the above yields that if $|A|_s \leq |\neg A|_s$ then $|A|_s \leq |Q|$. But then (a) yields that if $|A|_s \leq |\neg A|_s$ then $|DA|_s \leq |DQ| = 0$.

(∂b) $\partial 1 = 1$

(∂c) $\partial 0 = 0$, and if $0<v<1$, then $\partial v<v$

(∂d) If $v \leq v^*$ then $\partial v = 0$, where * is the operator that corresponds to negation.

This is stronger, because V_c might contain values that no sentence of the language could possess; the most that follows from (a)-(d) is that a *partial* operator ∂ *defined on the subset of* V_c *that is in the range of the assignment function* satisfies (a)-(d), but (e) adds that this operator is total. The stronger condition seems natural if not irresistible, and I have no objection to adding (e) as a requirement. Call an operator D satisfying these conditions a *determinacy operator* (and call the corresponding ∂ satisfying (∂a)-(∂d) a ∂*-operator*). If we require only (c_w) in addition to (a), (b) and (d), D will be called a *weak determinacy operator*; similarly, weakening (∂c) to

($\partial \mathbf{c}_w$) $\partial v \leq v$

gives the requirements on a *weak* ∂*-operator*.

Given a G-semantics for a language without a determinacy operator, can we extend it to one with such an operator? We certainly can extend it to one with a weak determinacy operator; and for all versions of G-semantics I know of, we can extend it to a full determinacy operator. The simplest way to show this is to explicitly define such an operator D from the connectives we already have in the language. One particular such D (I'll call it **D**) can be defined as

$$A \wedge \neg(A \rightarrow \neg A).$$[33]

Obviously it corresponds to an operator in the underlying space, viz. $\partial v =_{df} glb\{v, (v \Rightarrow v^*)^*\}$. That the conditions ($\partial$a), ($\partial$b), ($\partial c_w$) and ($\partial$d) are satisfied is apparent: e.g. to verify (∂d), we simply observe that if $|A|_s \leq |\neg A|_s$ then $|A \rightarrow \neg A|_s = 1$, so $|\neg(A \rightarrow \neg A)|_s = 0$, so $|A \wedge \neg(A \rightarrow \neg A)|_s = 0$. And the full ($\partial$c) is satisfied in the versions of G-semantics I'm familiar with, e.g. that of [5]. I'm mostly interested in G-semantics in which full determinacy operators are definable, but weak determinacy operators would in fact fit my main purposes, and will rely below on the fact that they are definable in *every* G-semantics.

An advantage of treating the determinacy operator as defined within the original language is that doing so settles the application of 'True' and 'Satisfies' to sentences containing this operator, and settles it in such a way that the biconditionals (T) and (S) and the corresponding intersubstitutivity theses are bound to hold even for sentences that involve the determinacy operator. Thus with 'determinately' defined as **D**, it is immediate that **there can be no new paradoxes of determinacy**.

[33] In [3] I used the alternative definition $A \wedge [\top \rightarrow A]$, where $\top$ is a logical truth; this is equivalent to the definition in the text in the particular G-logic considered there, but I believe that in G-logics for which the equivalence fails, the operator in the text is more useful.

Two examples that might initially be thought paradoxical are the "strengthened Liar sentence" Q_{-1} which directly or indirectly asserts that it is determinately not true (and hence must be a fixed point of the "strong negation operator" $\mathbf{D}\neg$), and the "weakened Liar sentence" Q_1 which directly or indirectly asserts that it is not determinately true (and hence must be a fixed point of the "weak negation operator" $\neg\mathbf{D}$). Let's focus on the latter. Q_1 is equivalent to $\neg\mathbf{D}[True(\langle Q_1\rangle)]$, and hence (by the Tarski biconditionals and substitutivity principles) to $\neg\mathbf{D}Q_1$, so

(*) $\quad |Q_1| = |\neg\mathbf{D}Q_1|$ (or equivalently, $|\neg Q_1| = |\mathbf{D}Q_1|$).

What can we say about the value of $\mathbf{D}Q_1$? First, $|\mathbf{D}Q_1|$ isn't 0: for if it were, then by (*), $|Q_1|$ would be 1, and we'd have a gross violation of (b). Second, $|\mathbf{D}Q_1|$ is strictly less than $|Q_1|$: for since $|\mathbf{D}Q_1| \neq 0$, we have by condition (d) that $|Q_1| \not\leq |\neg Q_1|$, so by (*) $|Q_1| \not\leq |\mathbf{D}Q_1|$; together with $|\mathbf{D}Q_1| \leq |Q_1|$ (condition (c_w)), this yields $|\mathbf{D}Q_1| < |Q_1|$. So application of the operator corresponding to $\mathbf{D}$ strictly lowers the value of Q_1, but doesn't reduce it to 0.

We do have, though, that a double application of this operator reduces the value to 0; that is, $|\mathbf{DD}Q_1| = 0$ (from which it follows that *it is not determinate that Q_1 is not defective*, i.e. $\neg\mathbf{D}\neg[\neg\mathbf{D}Q_1 \wedge \neg\mathbf{D}\neg Q_1]$[34]). The argument for $|\mathbf{DD}Q_1| = 0$ is that by condition (c_w), $|\mathbf{D}Q_1| \leq |Q_1|$; so by (*), $|\mathbf{D}Q_1| \leq |\neg\mathbf{D}Q_1|$; and $|\mathbf{DD}Q_1| = 0$ then follows by condition (d). It follows, of course, that $\mathbf{DD}$ (the result of applying $\mathbf{D}$ twice) is strictly stronger than $\mathbf{D}$. (The same conclusion would have emerged from studying the "strengthened Liar": the fact that $\mathbf{DD}$ is strictly stronger than $\mathbf{D}$ shows up in the fact that $|\mathbf{D}\neg Q_{-1}| > 0$ but $|\mathbf{DD}\neg Q_{-1}| = 0$.)

Notice that this argument doesn't turn on the specific definition proposed for $\mathbf{D}$; it doesn't even turn on the fact that $\mathbf{D}$ is definable in terms of the other connectives. Rather, it turns only on the general features of a G-semantics plus conditions (b), (c_w) and (d); not only was (c) weakened to (c_w), but (a) wasn't used. Indeed, we didn't even use the full strength of (b), we only used

($\mathbf{b}_w$) If $|A|_{M,s} = 1$ then $|DA|_{M,s} > 0$,

(whose inferential analog is $A, \neg DA \vDash \perp$, where $\perp$ is an absurdity). (b_w), (c_w) and (d) are the conditions for what I'll call a *very weak determinately operator*. We see then that for any such operator in a language with a G-semantics, DD is strictly stronger than D. Calling an operator D *idempotent* if DD is the same as D, we get

> **Conclusion:** No operator in a language with a G-semantics that meets the conditions for being even a *very weak* determinately operator can be idempotent.[35]

[34] For this simplifies to $\neg\mathbf{D}[\mathbf{D}Q_1 \vee \mathbf{D}\neg Q_1]$, which by the nature of Q_1 is equivalent to $\neg\mathbf{D}[\mathbf{D}Q_1 \vee \mathbf{DD}Q_1]$, which by two successive applications of $\neg\mathbf{DD}Q_1$ is valid.

We can't assert that Q_1 is defective (or, of course, that it isn't); only that it is "possibly" defective, i.e. not determinately non-defective.

[35] Also, if D meets those conditions then $D\neg D$ is strictly stronger than $\neg D$: for $|D\neg DQ_1|$ is $|DQ_1|$ and $|\neg DQ_1|$ is $|Q_1|$, and we've already seen that $|DQ_1| < |Q_1|$.

(It was implicit in our result on "truth-like operators" in Section 3 that no very weak determinately operator in a G-semantics can be bivalent, i.e. can be such that the value of any sentence of form DA is always 0 or 1; indeed, that result doesn't even require (d). It's also worth noting that for any operator that meets the full condition (c), idempotence immediately implies bivalence: (c) implies that for any A, if $|DDA|_s = |DA|_s$ as it must by idempotence, then $|DA|_s$ can only be 1 or 0. So (b_w) and the full (c) immediately rule out idempotence in any G-semantics.[36])

I must emphasize that these conclusions only apply to operators in languages with a full-fledged G-semantics; that is, languages with a DMC-semantics *for which the Tarski biconditionals hold*. Obviously if one gives up the latter assumption, the conclusion no longer need hold. For instance, suppose we were to stipulate that the original language is to be extended by the addition of an operator D^* which is stipulated to behave in the semantics as follows:

$|D^*A|_{M,s}$ is 0 if $|A|_{M,s}<1$, and $|D^*A|_{M,s}$ is 1 if $|A|_{M,s}$ is 1.

(D^* is of course just the operator version of the alleged "absolute designateness" predicate discussed in the previous section.) Such a D^* is by stipulation bivalent, and is easily seen to satisfy all of the conditions (a)-(d); our conclusions thus imply that adding it to the language would force a failure of the Tarski biconditionals, which we knew already (from the discussion of attempts at a model-theoretic revenge argument).

Might we contemplate settling for a determinacy operator that maintains idempotence and the Tarski biconditionals by weakening one of the conditions (b_w), (c_w) and (d)? One's first inclination might be to weaken (d), but as remarked in note 32, there is no way to do this in the best-known versions of G-semantics without either weakening (a) or giving up the idea that $|DQ| = 0$, either of which seems an intolerably high cost to pay. (And in *every* G-semantics, keeping idempotence together with (b_w) would rule out acceptance of the full (c).)

Another way to keep idempotence in a G-semantics would be to say that the D operator has non-trivial application only to formulas that don't themselves contain D: the line would be that if A contains D, $|DA|_s = 0$, so DD is the 0 operator and hence trivially idempotent.[37] This would allow us to preserve (c) and (d), and it generates a restriction on (b_w) that blocks the above proof (and it generates a restriction on (a) as well): in particular, the various "determinate Liar" sentences Q_α ($\alpha \geq 1$) now all get value 1. But limiting the scope of the determinacy operator in this way seems to me a very high cost. One could ameliorate that cost by introducing a hierarchy of determinacy operators, e.g.

[36] In inferential form, (c) amounts to the law $B \rightarrow DB \vDash B \vee \neg B$, which implies $DB \rightarrow DDB \vDash DB \vee \neg DB$. Idempotence would establish the premise, so we'd get excluded middle for D claims, which would suffice for inferential paradox.

[37] Of course, this will have to affect the application of the D operator to some sentences that don't contain 'D' but contain 'True' or 'Satisfies', if the schemas (T) and (S) are to be preserved.

an operator D_2 that applies non-trivially to formulas containing the "ground level" determinacy operator D; but D_2 couldn't be explained as DD since DD would just be D. It seems to me that introducing such a hierarchy of operators would throw away whatever virtues idempotence might be thought to have, and that the idea of a hierarchy of *primitive* determinacy operators has far less appeal than a hierarchy obtained by iterating a single determinacy operator.

The fact that the G-semantics rules out the existence of an idempotent operator satisfying the conditions of a determinacy operator (or even a *very weak* determinacy operator) would lead to a revenge problem if we had reason to believe that such an operator was intelligible: it would show a serious expressive limitation in any such language. But I maintain that there is no good reason to think that idempotent determinacy operators, or even idempotent very weak determinacy operators, are intelligible. I've already considered one argument for the intelligibility of the particular idempotent operator D^*, and argued that this argument rests on a misunderstanding of the significance of the model theory. But the possibility that there are more sophisticated arguments for idempotence remains to be considered.

11. Iterations of Determinacy Operators. The impossibility of idempotent operators in a G-semantics is a more stringent ban that might at first appear. To see this, observe that if D is any determinacy operator then so is DD (the result of applying it twice); that is, DD satisfies (a)-(d) if D does. Similarly, if D is any *weak* determinacy operator then so is DD. (Not so for *very* weak.) Because of this, our conclusion that if D is a weak determinacy operator in a G-semantics then it can't be idempotent can be extended: not only can't D be idempotent, DD can't either. That is, DD can't be identical to $DDDD$; from which it easily follows that it can't be identical to DDD. Continuing in this vein, we can argue that as n increases, the result D^n of applying D n times becomes strictly stronger as n increases.

We can also extend the iteration process a good way into the transfinite, as I will discuss in detail in Part Four. The basic idea, restricted for simplicity to the case where A is a sentence, is that for each limit ordinal λ for which the iteration is defined, $D^\lambda A$ says that for all $\alpha<\lambda$, $D^\alpha A$ is true; as a result,

$$|D^\lambda A| \text{ is the greatest lower bound of } \{|D^\alpha A| \mid \alpha<\lambda\}.$$

More generally, we can get that even when A can contain free variables,

(LIM) $\quad |D^\lambda A|_s$ is the greatest lower bound of $\{|D^\alpha A|_s \mid \alpha<\lambda\}$.

It's then easy to verify that as long as the iteration process is defined in accordance with (LIM) at each stage, it always leads from determinacy operators to determinacy operators and from weak determinacy operators to weak determinacy operators. So the anti-idempotence result shows that **we get a hierarchy of operators that become strictly stronger for as long as the iteration is satisfactorily definable** (i.e., definable in a way that accords with (LIM)). The limits on how far it is so definable raise interesting philosophical issues which I will discuss in Part Four.

This fact that the hierarchy never collapses to idempotence can be directly checked, by producing a hierarchy of "increasingly paradoxical" sentences. The simplest such hierarchies are the **transfinite Liar hierarchies**. There are two of them, "going in opposite directions"; I will focus on the fixed points Q_α of the "increasingly weak negations" $\neg$, $\neg D$,$\neg DD$, ..., but I could just as well have used the fixed points $Q_{-\alpha}$ of the "increasingly strong negations" $\neg$, $D\neg$, $DD\neg$, ...[38] So: Suppose we have defined D^α for some ordinal α. Then we can find an "α-level weakened Liar sentence" Q_α which says $\neg D^\alpha True(\langle Q_\alpha \rangle)$. By the Tarski biconditionals and substitutivity principles (which by definition hold in any G-semantics), it follows that

(**) $\quad |Q_\alpha| = |\neg D^\alpha Q_\alpha|$.

$|D^\alpha Q_\alpha|$ isn't 0: for if it were, then by (**), $|Q_\alpha|$ would be 1, and we can easily see (using (b) and (LIM)) that then $|D^\sigma Q_\alpha|$ would have to be 1 for all σ, contradicting our assumption. However, $|D^\alpha Q_\alpha| \leq |Q_\alpha|$ (by (c_w)); so by (**) $|D^\alpha Q_\alpha| \leq |\neg D^\alpha Q_\alpha|$; so by (d), $|D^{\alpha+1} Q_\alpha| = 0$. So $D^{\alpha+1}$ is a strictly stronger operator than D^α (from which it follows that it is not $D^{\alpha+1}$-true that Q_α is non-defective[39]).

To summarize, the situation is that for a fixed Q_α, the $D^\sigma Q_\alpha$ get stronger and stronger as σ increases, until σ reaches $\alpha + 1$; after that point no strengthening of the sentence by adding 'determinately' is possible, since the sentence already has value 0. But there are other sentences, e.g. the Q_β for $\beta > \alpha$, for which the iteration can proceed farther before collapsing to value 0; so the operators D^σ *never* collapse to the operator D^* or to any other idempotent operator.

The claim that the determinacy operator D, and the corresponding operator ∂ on the space of values, never collapses into idempotence may seem as if it couldn't be true: after all, for any partially ordered set V_c there is a cardinal d greater than the cardinality of all V_c-chains (linearly ordered subsets of V_c). But then for any formula A and assignment function s, the chain $\{|D^\alpha A|_s\}$ has stopped decreasing prior to the initial ordinal for d; so letting ξ be that initial ordinal, it seems that D^ξ must be idempotent.

The problem with this argument is that the iteration of the D operator breaks down (becomes undefined) before reaching cardinality d. There are two reasons why this is so.

The first reason for the breakdown depends on the nature of each value space V_c (used for models of cardinality no greater than c). In order to ensure that quantified formulas get values in the space, I required that V_c be c-complete: that is, each subset of V_c of cardinality no greater than c must contain a least

[38] Each $Q_{-\alpha}$ behaves very much like the negation of the corresponding Q_α.

[39] Indeed, for finite α, the 'α+1' can be replaced by 'α'. The proof of the parenthetical claim is a generalization of that in note 34: to say that it is not D^σ-true that Q_α is non-defective is to say that $\neg D^\sigma \neg [\neg DQ_\alpha \wedge \neg D \neg Q_\alpha]$, i.e. $\neg D^\sigma [DQ_\alpha \vee D \neg Q_\alpha]$, i.e. $\neg D^\sigma [DQ_\alpha \vee DD^\alpha Q_\alpha]$. This reduces to $\neg D^{1+\sigma} Q_\alpha$ (using the fact that $\neg D^{\alpha+1} Q_\alpha$), and this is valid whenever $1 + \sigma \geq \alpha + 1$; i.e. when $\sigma \geq \alpha + 1$ and α is infinite, or $\sigma \geq \alpha$ and α is finite.

upper bound (or equivalently, each such subset must contain a greatest lower bound).[40] I also gave reasons (note 19) for expecting that the space would not contain least upper bounds (or equivalently, greatest lower bounds) for all sets of higher cardinality than c. But the only acceptable way to define ∂^λ for a limit ordinal λ would be to let $\partial^\lambda(v)$ be the greatest lower bound of all $\partial^\alpha(v)$ for $\alpha<\lambda$; so if λ has cardinality greater than c, there is no reason to think that ∂^λ is defined.[41] Indeed, the fact that we have shown all iterations of ∂ to be idempotent, together with the result of two paragraphs back that if ∂^ξ were defined it would be idempotent, shows that there must be subsets of V_c of cardinality no greater than that of ξ that have no greatest lower bounds.

But there is a second reason for the breakdown in the iteration of the D-operator, which occurs much earlier and is of more interest for generating revenge problems. It arises from the fact that the language L (like all languages, in any but a special technical sense of 'language') can contain only countably many expressions. From this it follows that there are *countable* α for which there is no operator D^α in the language; this is a much earlier breakdown in the iteration process than the one required by the previous paragraph. This earlier breakdown is a source of revenge worries, and in Part Four I will investigate in detail how it occurs.

So for both of these reasons, the iteration process breaks down, and that is how the claim that it never collapses to idempotence avoids absurdity. And to repeat, it not only avoids absurdity, we have proved that it *must* hold given the basic assumptions of G-semantics and the very weak conditions (b_w), (c_w) and (d). Of course, one might hold on to idempotence and those weak conditions (or their inferential versions), if one were to give up the conditions on a G-semantics: for instance, if one were to give up the truth schema. But that has high costs. The main goal of the rest of the paper is to argue that giving up on idempotence does not have an intolerably high cost, and indeed is very natural.

12. Non-idempotent Determinacy versus Stratified Truth. One might think that a hierarchy of non-idempotent iterations D^α of a determinacy operator D would give rise to all the problems of stratified theories of truth and satisfaction in classical logic. But there are at least three reasons why this is not so (of which I take the first and third to be most important).

The first is that determinacy is a much more peripheral notion than truth, and we do have a unified notion of truth (and of satisfaction too). It is the notions of truth and satisfaction, not of determinate truth, that we need to use as devices of generalization. We've seen that stratifying the truth predicate

[40] The greatest lower bound of S is $(\sqcup\{v^* | v \in S\})^*$, where * is the operator on V corresponding to negation.

[41] Well, we could define $\partial^\lambda(v)$ as the greatest lower bound of all $\partial^\alpha(v)$ *for* α *in a sequence that is cofinal with* λ; but then if λ has no cofinal sequence of cardinality less than or equal to c, there is no reason to think that ∂^λ is defined. And the first ordinal of cardinality greater than c has no cofinal sequence of cardinality less than or equal to c, so this liberalization takes us no further.

seriously cripples our ability to make generalizations; not so for a "stratification" of the notion of determinacy.

The second point to make is that there is a serious disanalogy between the stratification of truth in classical theories and the "stratification" of determinate truth here. For in this theory, all the determinacy predicates are defined by iterating a single determinacy operator (and using a truth predicate); whereas in the case of classical stratified truth theories, each truth predicate must be introduced separately.

The third point is that we can reasonably hope, for each α, that our overall theory of truth and satisfaction and determinacy is $D^{\alpha}True$. This is in marked contrast to the classical truth theory case, according to which e.g. '$True_{\alpha}(\langle A\rangle) \rightarrow A$' is an important part of the theory but not true_{α} (but only $\text{true}_{\alpha+1}$). Thus the main objection that I raised against stratified theories (the one with which I closed Section 2) simply doesn't arise against the current theory.

Despite these three points, it may seem counterintuitive to suppose that there is no intelligible notion of a sentence being $True$ and $DTrue$ and D^2True and ..., through all iterations of the determinately operator; and that is what my account implies. Indeed it may seem not merely *counterintuitive*, it may seem *incompatible with point one*, i.e. incompatible with the point that we can use 'True' to make generalizations that one couldn't make otherwise. These two qualms are connected: one can't fully remove the intuitive pull of the first qualm without coming to understand why the second qualm is incorrect. And showing that the second qualm is incorrect requires the more precise treatment of the hierarchies that is to be given in Part Four. Even so, it is worth making a preliminary remark about the second qualm, and then addressing the first.

The incompatibility qualm: As I have been discussing the determinacy operators *so far*, the ordinal superscript 'α' in 'D^{α}' is not a variable; the intent has been rather that for each ordinal α in a suitable segment of the ordinals, there is a corresponding operator 'D^{α}' in the language. But (1) it looks as if we will be able to use the truth predicate to *get the effect of* quantifying over the α, even if the superscript isn't a variable. In particular, it looks as if we can express such thoughts as that for each α, the result of prefixing A with the α^{th} iteration of 'D' is true. But (2) that would be in effect just the application to A of an operator "$D^{\alpha}True$ for all α", and that looks like an idempotent operator that is the infinite conjunction of all the $D^{\alpha}True$. I've argued against there being such an operator in the language (and indeed will be arguing that there is no good reason to suppose that such an operator is intelligible); so what gives?

The long answer to this question will be given in Part Four. The very short answer is that we can indeed introduce a hierarchy of iterations D^{α} of D *for variable* α, and hence allow quantification over the α; however, in order to make the quantification well-behaved we must impose a bound on the α, and there

are compelling reasons why there can be no unique such choice of bound.[42] Any bound we impose to keep the hierarchy of operators well-behaved can be relaxed, so that there is no maximal good bound. *Given any reasonable choice of bound for a hierarchy of iterations D^α of D*, we can then use the truth predicate to achieve what is in effect a quantification over the α, just as in the previous paragraph; but since the α are bounded, this will simply be another iteration of D in an enlarged hierarchy, so it does not achieve the intended effect.

The counterintuitiveness qualm is that it just seems as if we have a unified notion of hyper-determinate truth ("determinate truth in every reasonable sense of that term") corresponding to "$True$ and $DTrue$ and D^2True and ...". Or if you like, a unified notion of "defective in some reasonable sense of that term", viz. "($\neg True$ and $\neg False$) or ($\neg DTrue$ and $\neg DFalse$) or ($\neg D^2True$ and $\neg D^2False$) or ...".[43]

I don't want to deny that we have these notions; but not every notion we have is ultimately intelligible when examined closely. A large part of the response to the counterintuitiveness qualm will be an argument, in Part Four, that the notion of "the" hierarchy of iterations of D has a kind of inherent vagueness that casts doubt on there being a well-behaved notion of "D^α-true for every α"; and without that there is no reason to suppose that there is a well-behaved notion of "determinately true in every reasonable sense of that term". The apparent clarity of such notions is an illusion. (One can, to be sure, give *ill-behaved* definitions, that would seem well-behaved *if the inherent vagueness in the hierarchy were not taken account of*; so part of the response to the qualm will be to show how without faulty assumptions those definitions are indeed ill-behaved.)

A unified notion of hyperdeterminate truth, then, is basically something we should abandon. For this recommendation to be acceptable, following it had better not have the high intuitive costs that stratified truth theories have. And it doesn't. In addition to the three points made earlier in this section, I note the following. A substantial part of the counterintuitiveness of stratified truth theories stems from the fact that if such a theory were actually in use, each person would be under constant pressure to employ very high ordinal subscripts in order to ensure that what they said had sufficient strength; and because of ignorance about the subscripts employed by others whose views we are discussing, we would often end up employing too low a subscript to capture what we wanted to say. This is a point well illustrated by Kripke's discussion of Nixon

[42] A slightly longer answer is that we can introduce a hierarchy of pseudo-iterations D^α of D for variable α, which we can quantify over unrestrictedly; but for large α these can't be viewed as genuine iterations of D, nor will they be determinacy operators in any reasonable sense; and there is no satisfactory way to restrict to "good" ordinals (for which the D^α behaves as it is supposed to) except by restricting too far. (A still longer answer involves the idea that the notion of a "good iteration" is a "fuzzy notion" for which classical laws fail; this is why it is impossible to achieve satisfactory results by quantifying over only the good iterations.)

[43] Or alternatively, as "either $defective$ or $\neg D\neg defective$ or $\neg D^2\neg defective$ or ...", where '$defective$' means 'neither determinately true nor determinately false'. In some G-logics this is a slightly stronger predicate than the one in the text.

and Dean (mentioned at the end of Section 2); one of the symptoms is that if Nixon says "Everything Dean says is untrue$_\alpha$" and Dean says the corresponding thing about Nixon but with a possibly different subscript, then at least one of them fails to include the other's remark within the scope of his own. The situation with iterations of the determinacy operator is quite different: e.g. if Nixon says "Everything Dean says is D^α not true",[44] and Dean says the corresponding thing about Nixon, then both succeed in disagreeing with the other's remark even though they have used the same ordinal α. Because of this, there is little pressure to employ high ordinal superscripts on determinacy operators in normal contexts. And because of that, it is difficult to find circumstances where it is plausible to maintain that we should have reason to think that a person's theory is "defective in some sense of that term" without there being a sufficiently high α—say $\in_0$, or the first non-recursive ordinal, or even higher—for which the operator D^α is perfectly clear and for which we are in a position to think the theory to be not $D^\alpha True$.

Part Four: Transcending the Hierarchies?

13. A New Revenge Worry, in Three Strengths. Part Four will in effect be concerned with the question (raised in the last section) of why we can't use the truth predicate to "unify" the various iterations of the determinacy operator. This is closely related to a revenge worry: for if we *could* use the truth predicate to "unify" the various iterations of the determinacy operator to get a "hyperdeterminately" operator D_{hyp}, then it looks as if we could use that unified operator to produce a new "Hyper-Liar" paradox (via a sentence that asserts its own lack of hyper-determinate truth). Such a paradox couldn't be handled along the lines used for the paradoxes in the hierarchy $\{Q_\alpha\}$, since that solution depends on non-idempotence; and perhaps it couldn't be consistently handled at all without giving up the truth and satisfaction schemas. Of course, from known consistency results, there can't be a "unification" in the language that has this last feature; but it is prima facie puzzling why we can't use the truth predicate to create one.

But the concern in Part Four will not be only with the idea of hyperdeterminately operators that are *definable in the language*; the real worry is that the hierarchy of determinately operators used in working out a G-solution can be used to *make such a hyper-determinately operator intelligible*. We may not be able to define such an operator in the language, because of certain expressive limitations of the language; but the idea is that we can transcend these limitations in the mind ("mentally quantifying" over the levels of the hierarchy), and then contemplate *adding such an operator to the language*. The **strong revenge worry** is that adding such an operator to the language would produce

[44] I focus on the case of 'D^α not true' rather than 'not D^α true' to maintain the parallel with 'not true$_\alpha$' in the stratified truth theory. If $\alpha<\beta$, 'true$_\beta$' is weaker than 'true$_\alpha$' but 'D^β' is stronger than 'D^α'; so 'not true$_\alpha$' gets stronger as α increases, whereas 'not D^α' would get weaker. (However, the immediate claim in the text would hold just as well for 'not D^α true'.)

a new paradox that requires giving up the truth schema. Substantiating this would be a fatal blow to any claim that a G-solution adequately resolves all the paradoxes.

There are weaker revenge worries also based on the idea that we can use the hierarchy of determinately operators to make intelligible some sort of hyper-determinately operator. The **intermediate-strength revenge worry** is that such an operator can't be made consistent with the truth schema *in the semantic framework so far introduced*, that is, in a G-semantics. If that were right it would mean that if we were to expand the language to include these new concepts, we would get new paradoxes *that can't be resolved by the same sort of means by which paradoxes were resolved in the language L that has been treated* (though they might be resolved by other means). That too would undermine any claim that G-solutions offer an ultimate answer to the paradoxes. The natural way to try to argue for the intermediate-strength revenge worry is to argue that we can make intelligible an operator that is both idempotent and satisfies the inferential versions of the conditions on being a very weak determinacy operator (see Section 10); for we've already seen that such an operator couldn't possibly be treated within a G-semantics.

The **weak revenge worry**—so weak that maybe it shouldn't count as a revenge worry at all—is that we can use the hierarchy of determinately operators to make intelligible some sort of hyper-determinately operator not definable in the original language, which breeds new prima facie paradoxes. It is *not* claimed that these new paradoxes aren't resolvable in a G-semantics—that would be the intermediate-strength worry. So it wouldn't really undermine the claim that we need nothing more than G-semantics to resolve the paradoxes. Still, if the weak worry could be substantiated it would show that we couldn't make do with a G-semantics for a single language: a G-solution for a single language would generate a richer language that needs its own G-solution, and so forth. That wouldn't defeat the basic idea of G-solutions, but it would be a disappointment. It would show that *one* of the disadvantages of classical stratified truth theories carries over to G-solutions. (G-solutions would however still retain the first and third advantages discussed in the previous section; indeed, it is arguable that even the second would not be *totally* undermined.[45])

As I've noted, the only obvious way to try to argue for the intermediate-strength worry is to argue for the intelligibility of an *idempotent* determinacy operator (or at least, an idempotent "very weak determinacy operator"). For the weak revenge worry, this is not so: we could substantiate it by quantifying over all iterations of D *that are expressible in the language* L, and there's no obvious reason why the result of so doing should be idempotent. If it isn't idempotent, then in an expanded language L^* that includes it, we'd get a new hierarchy obtained by iterating D_{hyp}; that's how a G-semantics for L^* might

[45] Part of the reason is implicit in the point made at the very end of the previous section: there is normally little practical need to ascend very high in the iterations of D, in strong contrast to the case of stratified classical theories.

be possible (and hence why a substantiation of the weak worry would only be a relatively minor blow to G-solutions).

The prima facie case for a D_{hyp} operator being idempotent (and thus supporting at least the intermediate-strength worry) seems slight:

(i) We would get idempotence from the assumption that hyper-determinacy claims obey excluded middle (together with conditions (b_w) and (c_w) of Section 10). But it isn't evident how excluded middle for hyper-determinacy claims could be argued, short of either the assumption that excluded middle holds generally (which would of course rule out G-solutions from the start, independent of revenge worries) or the assumption that we can read a hyper-determinacy claim off the model theory (an assumption that I hope to have disposed of in section 9).

(ii) We could plausibly get idempotence from the assumption that a hyper-determinacy predicate would unify *all* iterations of D, *even those not expressible in the language* L. In particular, suppose that one of the "iterations of D" included in the "unification" would be "hyper-determinately hyper-determinately"; then "hyper-determinately" would have the full-strength of "hyper-determinately hyper-determinately", and so (assuming condition (c_w) of Section 10) "hyper-determinately" would be idempotent. But there seems little basis for the thought that we can define a hyper-determinateness operator that unifies *even those iterations of* D *not expressible in the language* L. (Perhaps we could get such an argument from the assumption that 'is an iteration of D' is a bivalent predicate; but we'll see that that assumption is wholly unwarranted.)

I might add that even if there were an idempotent hyper-determinacy operator, that wouldn't seem to support the strong revenge worry: presumably one could avoid paradox in various ways that are consistent with keeping the truth schema, e.g. by denying the iterability of the operator or by in some other way restricting the inference from A to $D_{hyp}A$. Admittedly, such solutions are unattractive; and my view is that the rationale for a G-solution would be thoroughly undermined if it could be argued that an idempotent determinacy operator is intelligible.

As I've said, my goal is in fact stronger: I will argue that there is no basis for *even the weak form* of the worry; from which it follows of course that there is no basis for the intermediate or strong forms. But a word of clarification is in order. It is certainly not part of my claim that it is incoherent to imagine that the language L be expanded in non-definitional ways. There is always a possibility of introducing new concepts; this is certainly the case for concepts pertaining to new forms of society or new organisms or newly discovered particles, and I see no reason to doubt that it is so for new mathematical concepts as well. And it may well be that adding new mathematical concepts to the language could make further iterations of the determinacy operator expressible in the language, and consequently would lead to an extension of the G-solution to the enriched language. (There would be a completely mechanical way of making the extension.) I take it that *that* wouldn't count as any kind of problem for a

solution to the paradoxes, and so I'm understanding the "weak revenge worry" to require more than this. What it requires to substantiate the weak revenge worry is that *simply by reflecting on the hierarchy of determinacy operators* we can make intelligible an operator that transcends them; it is that, and not merely that we might make this intelligible in some other way, that seems to generate "levels of language" in some objectionable sense. This is vague—the weak form of the revenge worry *just is* vague.[46] But I think that what follows will undermine the worry, and in the process undermine the intermediate and strong worries too.

14. Hierarchies of Operators: Introductory Remarks. In order to properly discuss this, we need to be much clearer about transfinite hierarchies of iterations of a determinacy operator.

Any iteration D^α of D will be a syntactic operator that takes any L-formula A to an L-formula $D^\alpha A$ with the same free variables.[47] For any formula A, we can take $D^0 A$ to just be A; that is, we can take D^0 to simply be the identity operator on L-formulas. If we've defined the operator D^α, then we can define the operator $D^{\alpha+1}$: for any formula A, the result of applying $D^{\alpha+1}$ to A is to be the result of applying D to the result of applying D^α to A.

So the only issue in specifying the hierarchy of iterations of D is specifying D^λ for limit λ. Of course, we want to do this in such a way that for any formula A, $D^\lambda A$ is in effect the infinite conjunction of the $D^\alpha A$ for $\alpha<\lambda$. There will be a limitation on how far we can do this, so let us say that we want it for all limit ordinals λ less than a certain limit ordinal σ; σ will then be called the length of the hierarchy.

More precisely, let OP be the set of operations on formulas of L that assign to each formula another formula with the same free variables; and for any $O \in OP$, let $det(O)$ be the operation that takes each formula x into the result of applying D to Ox.

> **Definition:** A *hierarchy (of iterations of D)* is a function H from $\{\alpha \mid \alpha < \sigma_H\}$ to OP, for some limit ordinal σ_H, that meets the following conditions:

[46] The intermediate and strong forms can be freed of similar vagueness, for they can be taken to involve intelligible expansions of the language however achieved.

[47] Each D^α is a recursive operator, since once its application to a specific formula such as '0 = 0' is specified, there is a mechanical procedure for turning that into a specification of its application to any other formula A. (The mechanical procedure is intuitively a kind of "generalized substitution": it involves not only substituting A for appropriate occurrences of '0 = 0', but also substituting the standard name of A for appropriate occurrences of the standard name of '0 = 0', the standard name of the standard name of A for appropriate occurrences of the standard name of the standard name of '0 = 0', and so on.) Since each such operator is recursive, it is definable in the 'true'-free fragment L_0 of L. (The word 'true' will occur in sentences that are *mentioned* in the definitions of D^λ for limit λ, but that is just syntax and doesn't prevent the definition being in L_0.) Of course, the fact that each specific D^α is definable in L_0 doesn't show that the whole hierarchy of them is; and the hierarchies I will focus on will turn out to be far too large to be definable in L_0.

RCZ $H(0)$ is the identity operator;

RCS For any $\alpha<\sigma$, $H(\alpha+1)$ is $det(H(\alpha))$.

RCL For any limit ordinal $\lambda<\sigma_H$, $H(\lambda)$ is a member of OP such that for any L-formula x and any assignment s of objects to the free variables of x, s satisfies $[H(\lambda)](x)$ if and only if for all $\beta < \lambda$, s satisfies $[H(\beta)](x)$.

I will call these the *Reasonability Conditions* for the behavior of a hierarchy on zero, successors, and limits respectively. The condition (LIM) of Section 10 was just the model-theoretic analog of (RCL).

A minor complication here is that RCL uses 'satisfies' in a way slightly differently from the way used so far: it speaks of satisfaction of formulas *with arbitrarily many free variables* by *assignments of objects to the free variables*, whereas I have taken satisfaction to be of formulas *with a single free variable* by *objects*. There is nothing deep here: by a slight extension of the ideas in note 4, the use of 'satisfies' employed in (RCL) could be defined from my official one. (I spare you the details.)

Of course, 'satisfies' in this modified sense is still a non-classical notion: excluded middle cannot be assumed to hold generally for it. This gives the condition (RCL) a rather different character than conditions (RCZ) and (RCS). And it raises a point which will turn out to have major significance: **we have no obvious reason to think that 'is a hierarchy' is a predicate that obeys excluded middle.** Indeed, we'll see that the supposition that it obeys excluded middle leads to contradictions. However, we'll also see that we can define more restrictive notions of hierarchy for which excluded middle can be assumed.

Although this definition allows for a multiplicity of hierarchies (not only hierarchies of different lengths, but also different ones of the same length), it is *roughly* the case that different hierarchies of the same length are "equivalent", and that those generated by paths of different lengths are "compatible". More precisely, call two operators O_1 and O_2 on L-formulas *equivalent* if for every L-formula A and every assignment function s, s satisfies O_1A iff it satisfies O_2A. Call two hierarchies H_1 and H_2 *compatible* if for every ordinal α in the domain of both, $H_1(\alpha)$ is equivalent to $H_2(\alpha)$. (And call two hierarchies *equivalent* if they are compatible and have the same length.) Then we have the following:

> **Equivalence Theorem:** H_1 and H_2 are hierarchies $\vDash$ H_1 is compatible with H_2.

Note that I have not stated this as a conditional, and can't do so in general because of the problem with excluded middle. (It is easy to see that in G-logics, $\rightarrow$-introduction is only valid on the assumption of excluded middle for the premise.)[48] The equivalence claim in the first sentence of this paragraph

[48] All of the premises, if there are side formulas.

did state it as a conditional, which is why I emphasized that it was only a rough approximation to the truth.

Proof of Equivalence Theorem: an obvious induction, but it is worth spelling out to ensure that no fallacious use is made of excluded middle. So, assume that H_1 and H_2 are hierarchies, A is any formula, and s is any assignment function. We need that for all $\alpha < \min\{\sigma_{H_1}, \sigma_{H_2}\}$, s satisfies $[H_1(\alpha)]A$ if and only if it satisfies $[H_2(\alpha)]A$. By the transfinite induction rule, which is valid even for non-classical predicates (see end of Section 5), it suffices to show that for any $\alpha < \min\{\sigma_{H_1}, \sigma_{H_2}\}$,

> $(\forall\beta < \alpha)(s$ satisfies $[H_1(\beta)]A$ if and only if s satisfies $[H_2(\beta)]A) \rightarrow s$ satisfies $[H_1(\alpha)]A$ if and only if s satisfies $[H_2(\alpha)]A$.

Given the supposition that H_1 and H_2 are hierarchies, the conclusion is evident when α is 0 or a successor. For limit λ, the supposition that they are hierarchies gives that for any $\alpha < \min\{\sigma_{H_1}, \sigma_{H_2}\}$,

> $[s$ satisfies $[H_1(\alpha)]A$ if and only if $(\forall\beta < \alpha)(s$ satisfies $[H_1(\beta)]A)] \wedge [s$ satisfies $[H_2(\alpha)]A$ if and only if $(\forall\beta < \alpha)(s$ satisfies $[H_2(\beta)]A)]$,

which yields the desired conclusion via the inference $(X_1 \leftrightarrow Y_1) \wedge (X_2 \leftrightarrow Y_2) \vDash (Y_1 \leftrightarrow Y_2) \rightarrow (X_1 \leftrightarrow X_2)$, which is easily seen to hold in all G-logics by several applications of Condition (II) from Section 10. ■

A crucial question is, how far can we get hierarchies to extend? It's easy enough to satisfy the instances of (RCL) when λ is sufficiently simple: e.g., when λ is ω. When A is a formula with a single free variable 'v', we could let $D^\omega A$ be the formula

> For all n, v satisfies every formula that results by prefixing $\langle A\rangle$ with n occurrences of D.

This has the same free variable that A does. In addition to its occurring freely, a formula that contains it is mentioned, but this is unproblematic. (And we can remove the restriction to formulas with only 'v' free: e.g. let $D^\omega A$ be

> For all n, every formula that results by prefixing $\langle A\rangle$ with n occurrences of D is satisfied by every assignment s of v_{i_1} to $\langle v_{i_1}\rangle$ and ... and v_{i_k} to $\langle v_{i_k}\rangle$;

where these are the free variables of A.)

So specifying D^ω is easy; but as the limit ordinal λ becomes more complicated, specifying D^λ becomes more difficult. Indeed it can't be done at all when λ is sufficiently large, for instance, when λ isn't definable in L or when any of its predecessors isn't definable in L.

Moreover, when it *can* be done, there will often be significantly different ways to do it, corresponding to different ways that λ and its predecessors might be defined. In particular, we might define λ as the supremum of a certain sequence

S of ordinals; and then we might be able to take $D^\lambda A$ as saying roughly that for all ordinals α in S, the result of prefixing an appropriate operator D^α to A is true (or true of v, if A contains 'v' free). But if there is one sequence with λ as its supremum there will be many, so the precise choice of the operator D^λ will not be unique. And the non-uniqueness increases as λ increases, because then many of the D^α from which D^λ is defined will themselves not be unique. In short, a specification of D^λ that takes the form suggested will depend on a whole path p of definitions of limit ordinals. Indeed, we'll see that it is possible to define a function Φ that, roughly speaking, "takes paths to hierarchies": given that p is a path of definitions of limit ordinals, $\Phi(p)$ will be a hierarchy. I'll sometimes call it D_p, to emphasize that it is a path-dependent hierarchy of iterations of D. (I'll also use the notation ${D_p}^\alpha$ as a suggestive abbreviation for $[\Phi(p)](\alpha)$.) (I will give a more rigorous treatment of paths and path-dependent hierarchies in ensuing sections.)

There are many paths of a given length if there are any at all, and the hierarchies generated by distinct paths are distinct. It would be possible to live with this high degree of non-uniqueness of the hierarchies, given the equivalence theorem. But it is more convenient to restore as much uniqueness as we can. An obvious idea for doing so is to existentially quantify over the paths, i.e. to let $D^\alpha A$ be defined as something like '$\exists p[Path(p) \wedge \langle {D_p}^\alpha A\rangle$ is true]'. We'll see that as long as we put certain bounds on the length of the paths, or equivalently on the ordinal α, it is possible to fully restore uniqueness by such a route. But when we lift those bounds, uniqueness can't be fully restored without making the definition virtually worthless, and this is crucial to the dissolution of revenge problems.

15. "Small" Hierarchies. I'll eventually want to consider hierarchies of iterations of D that extend "as far as possible" through the ordinals.[49] But the fact that the notion of a hierarchy cannot be assumed to obey excluded middle will complicate the discussion, and so in this section and the next I will give a "warm-up" that restricts to "small" hierarchies in which this problem does not arise. More specifically, let L_R be some fixed fragment of L for which we know excluded middle to hold. (It might for instance be the 'true'-free fragment L_0, or the fragment L_1 consisting of sentences in which 'true' occurs only in the context 'true and a sentence of L_0'.) Let λ_R be the first ordinal that is not definable in L_R. The path-dependent hierarchies to be discussed early in this section can have lengths up to and including λ_R, and by the end of the section I will have unified them into a single path-independent hierarchy of length λ_R. Even when the fragment of L under consideration is L_0, the hierarchies to be considered are thus very much larger than those considered in [3]: in that paper I imposed very stringent requirements on the hierarchies, which entailed that their length had to be a recursive ordinal. I can now see no good motivation for those stringent requirements.

[49] In the same sense that someone might want to be "as rich as possible", even if he didn't think a state of "maximal richness" made sense.

Let σ be any limit ordinal no greater than λ_R, i.e. any limit ordinal all of whose predecessors are L_R-definable. In particular, every *limit* ordinal less than σ is L_R-definable. (The latter claim is really no weaker than the former, since there is an obvious way of obtaining a definition of a successor ordinal from a definition of the largest limit ordinal that precedes it.) So there is a function p—many of them in fact—that assigns to each limit ordinal λ less than σ some L_R-definition of it, i.e. some L_R-formula (with exactly one free variable) that is satisfied by λ and by nothing else. Call any such function p an *L_R-path of length σ*. Note that σ is determined by p: given any L_R-path p, its "length" (in my slightly nonstandard sense) must be the first limit ordinal for which p is undefined, which I'll call σ_p. Also, note that 'L_R-path' involves the notion of L_R-definability and hence is not a term of L_R. But it is a term of the fragment L_{R^*} that results from L_R by allowing true' to occur in the context 'true and a sentence of L_R'; and excluded middle must hold throughout this fragment given that it does throughout L_R. So for any function p, either it is an L_R-path or it isn't.

To repeat, for any limit ordinal σ up to and including λ_R, there are L_R-paths p whose length is σ; and obviously this is not so for any $\sigma > \lambda_R$.

I now state a special case of a theorem proved in an Appendix to the paper. Let σ be any countable limit ordinal, let $Pred(\sigma)$ be the set of its predecessors, and $Pred_{\lim}(\sigma)$ be the set of limit ordinals that precede it. Let P_σ be the set of functions from $Pred_{\lim}(\sigma)$ to formulas with a single free variable 'μ' (which we can think of as a variable restricted to countable limit ordinals), and let P be the union of the P_σ for all countable σ. (P is thus an L_0-definable set to which we expect all L_R-paths to belong, whatever the fragment L_R.) Let OP be the set of operations on L-formulas, and let J be the set of functions from initial segments of the countable ordinals to OP. (J is thus an L_0-definable set to which we expect all hierarchies to belong.) Then:

> **Theorem on Existence of Small Hierarchies:** There is an L_0-definable function $\Phi : P \longrightarrow J$ such that for every L_R-path p, $\Phi(p)$ (aka D_p) is a hierarchy of length σ_p.

Since we have just seen that there are L_R-paths of any limit length up to and including λ_R, we have:

> **Corollary:** There are path-dependent hierarchies of any length up to and including λ_R.

Let an *L_R-hierarchy* be a hierarchy of form $\Phi(p)$ for p an L_R-path. An obvious but important fact about the notion of an L_R-hierarchy (for a specific L_R within which excluded middle holds) is that it obeys excluded middle: any function either is an L_R-hierarchy or isn't one. The theory of L_R-hierarchies can thus by developed without attending to the subtleties caused by possible failures of excluded middle. Note that an L_R-hierarchy needn't be *definable* in

L_R. Indeed, it needn't be definable even in the full language L, and it won't be if p itself is undefinable in L. One issue of some interest (though it won't be a central concern here) is: for which σ are there L_R-*definable* hierarchies of length σ. It is immediate that for a hierarchy (or a path) to be L_R-definable, its length σ must be *strictly less than* λ_R: for λ_R is by definition undefinable in L_R, but if a hierarchy (or a path) is L_R-definable then so is its length. It can easily be shown that there is no maximal length for L_R-definable hierarchies, indeed, each L_R-definable hierarchy has an L_R-definable proper extension.[50] I don't know if there are L_R-definable hierarchies on arbitrarily large proper initial segments of $Pred(\lambda_R)$, but even in the case of L_0 we have L_0-definable hierarchies extending well into the non-recursive ordinals.[51]

Even though specific L_R-hierarchies needn't be definable in the full L, we can nonetheless quantify over all of them (including the undefinable ones) in the "successor fragment" L_{R^*}. For sake of simplicity, I will define a hierarchy of operators that apply only to 'v'-formulas (formulas whose sole free variable is 'v'); this can be extended to a hierarchy of operators on arbitrary formulas by the route illustrated for D^ω in the previous section.

> **Definition of Path-Independent Hierarchy of Length λ_R:** If A is any 'v'-formula, let $D^\alpha_{[L_R]}A$ be the formula (whose free variables are 'α' and 'v')
>
> $\exists p(p$ is an L_R-path $\wedge\ \alpha < \sigma_p \wedge$ the result of applying $[\Phi(p)](\alpha)$ to $\langle A\rangle$ is true of $v)$.

(We could restrict the quantification to paths of length λ_R without affecting the result.) Since there are no L_R-paths of length greater than λ_R, $D^\alpha_{[L_R]}A$ is false of everything if $\alpha \geq \lambda_R$. Consequently, $D_{[L_R]}$ has useful application only when $\alpha < \lambda_R$; after this, it fails to meet condition (RCL) on being a hierarchy. But within this domain of useful application, $D_{[L_R]}$ behaves very nicely: for (i) using the Equivalence Theorem in the strong conditional form (which is legitimate in

[50] Let H be an L_R-definable hierarchy. If its length is a successor $\gamma+1$, extend it by adding $\langle \gamma+1, det(H(\gamma))\rangle$. If its length is a limit ordinal λ (which must be L_R-definable, since H is), extend it by adding $\langle \lambda, O^*_H\rangle$, where O^*_H is the operator that assigns to each formula x whose sole free variavle is 'v' the following formula:

> what results from substituting the L_R-definition of H and the standard name of x into the blanks in "For all operators O in the range of $\_\_$, v satisfies the result of applying O to $\_\_$)".

The definition of O^*_H really needs to be generalized to apply to arbitrary formulas, but one way to do that was illustrated in the discussion of D^ω in the previous section and another will be mentioned in the Appendix.

[51] For instance, let p be the function that assigns to the smallest non-recursive ordinal ν the formula 'μ is the smallest non-recursive ordinal', and assigns to each recursive ordinal the formula $j(\alpha)$^'is a Church-Kleene notation for μ', where $j(\alpha)$ is the *numerically smallest* Church-Kleene notation for it (in **O** or some other universal system). Then p is L_0-definable and is an L_0-path, with domain $\{\alpha \mid \alpha < \nu+1\}$. From this we could easily get L_0-paths extending much farther, e.g. to the limit of ν, ν^ν, ν^{ν^ν}, ... (and beyond).

contexts like this where we have excluded middle for the antecedent), we have that for each L_R-path p, the operator $D^\alpha_{[L_R]}$ is equivalent to D^α_p wherever the latter is defined; and (ii) for each $\alpha < \lambda_R$ there are L_R-paths for which it is defined. So we have a single quite natural hierarchy that usefully extends all the way up to λ_R.

16. A Length-Independent Hierarchy? Revenge? So far I've been holding the "effectively classical" fragment L_R fixed, and seeing what can be done within it. But as I've noted, whenever we have a fragment L_R that we know to be "effectively classical", we can enlarge it to a "successor fragment" L_{R^*}. The path-independent hierarchy $D_{[L_{R^*}]}$ usefully extends further than $D_{[L_R]}$ does: it extends up to λ_{R^*}, which is strictly larger than λ_R. In their common domain of usefulness, i.e. when $\alpha < \lambda_R$, the operators $D^\alpha_{[L_R]}$ and $D^\alpha_{[L_{R^*}]}$ will not be identical, for the formulas $D^\alpha_{[L_{R^*}]}A$ quantify over more paths. But the operators are nonetheless equivalent (when $\alpha < \lambda_R$): for any formula A, $D^\alpha_{[L_R]}A$ and $D^\alpha_{[L_{R^*}]}A$ are true of exactly the same things. In other words, the hierarchy $\{D^\alpha_{[L_{R^*}]} \mid \alpha < \lambda_{R^*}\}$ is *in effect* a proper extension of the hierarchy $\{D_{[L_R]} \mid \alpha < \lambda_R\}$ (though it doesn't *literally* extend it since it assigns different formulas at each infinite stage).[52]

In short, though we have achieved a certain kind of path-independence, we have not achieved length-independence: given any path-independent hierarchy of the sort described in this section, we can convert it to a longer one. We have a "hierarchy of path-independent hierarchies".

But isn't there a way to produce a unique hierarchy by "unifying" the ones we have? One might be tempted to argue as follows:

> Consider the set S of all λ_R for classical fragments L_R; there is a smallest limit ordinal ρ such that $\rho \geq \lambda_R$ for all λ_R in S. So for each $\alpha < \rho$, there are $\lambda_R \in S$ for which $\alpha < \lambda_R$; pick one, and let D^α be $D^\alpha_{[L_R]}$. This defines a hierarchy extending up to ρ which is guaranteed to be well-behaved (since at each stage it is equivalent to a well-behaved operator).

But this argument presupposes that it makes sense to speak of "the set of all (effectively) classical fragments L_R" (or rather, "the set of all λ_R for classical fragments L_R"; but the latter makes sense only if the former does). But that supposition is justified only if we can assume excluded middle for the predicate 'is a classical fragment'. And the assumption of excluded middle here is both *prima facie* unwarranted and demonstrably inconsistent.

It is *prima facie* unwarranted because to call a fragment effectively classical is to say that for each formula A within it, excluded middle holds. But we know from the end of Section 3 that we can't assume excluded middle for claims of

[52] Also, the two hierarchies don't even assign equivalent operators to ordinals that are outside the domain of useful application of one but inside the domain of useful application of the other.

form 'A obeys excluded middle': indeed, from excluded middle for 'A obeys excluded middle' one can infer excluded middle for A (note 14). If we can't assume excluded middle for all claims A, why should we be able to assume it for "L_R is a fragment all members of which obey excluded middle"?

But the key point is that it is actually *inconsistent* to assume excluded middle for the predicate 'is a classical fragment'. The reason for that is that *given that additional assumption* the argument displayed above is valid, and easily leads to the further conclusion that the "unified hierarchy" of length ρ is itself effectively classical. But then we can extend the hierarchy past ρ, by applying the Small Hierarchy-Existence Theorem to a path of length ρ. So it would follow that there is an effectively classical hierarchy bigger than all effectively classical hierarchies and so bigger than itself.

The fact that 'effectively classical fragment' is not a predicate for which excluded middle can be assumed makes it difficult to find useful generalizations of the form "For all effectively classical fragments ...". That is why in this section I have avoided doing so. Rather, I began the section by noting that L_0 is clearly an effectively classical fragment, as are L_1, L_2, and so forth; indeed, for each clear case L_R of such a fragment, its "successor fragment" L_{R^*} is one too. That is enough to give the neverending "hierarchy of path-independent hierarchies" that I have discussed.

Consequences for revenge? Suppose that we pick a particular path-independent hierarchy in the "hierarchy of hierarchies"; say $\{D_{[L_R]} \mid \alpha < \lambda_R\}$. Are we faced with even a weak form of revenge problem? Obviously not: we have defined ${D_{[L_R]}}^{\alpha}$ *for variable* α, so we can easily "unify" the operators ${D_{[L_R]}}^{\alpha}$ in *this* hierarchy simply by defining $D_{hyp(L_R)}A$ as

$$\forall\alpha(\alpha < \lambda_R \rightarrow {D_{[L_R]}}^{\alpha}A).$$

$D_{hyp(L_R)}$ is not a member of the hierarchy $\{{D_{[L_R]}}^{\alpha} \mid \alpha < \lambda_R\}$, but it is definable in L (and indeed, in L_{R^*}), by the definition just given; indeed, it is just the ${\lambda_R}^{th}$ member of the proper "extension" of that hierarchy $\{{D_{[L_{R^*}]}}^{\alpha} \mid \alpha < \lambda_{R^*}\}$. Since $D_{hyp(L_R)}$ is definable in L, the general consistency result applies to it: so we have a guarantee that $D_{hyp(L_R)}$ does not lead to paradox.

The point of this discussion is simply to serve as a warm-up for the discussion that follows of what can be done in the full L. Obviously the discussion so far does nothing to dispel the worries of Section 13: it simply shows that it is possible within the language to transcend the hierarchy of iterations D^{α} *for those* α *that are definable in a single demonstrably classical fragment of the language*. The results might even encourage the weak form of revenge worry: for the discussion shows that we can intelligibly transcend a hierarchy of iterations of ${D_{[L_R]}}^{\alpha}$ for α definable in L_R, but only by going to a higher language L_{R^*}; which might suggest that we can intelligibly transcend a hierarchy of all those iterations D^{α} of D that are definable in the full L, but only by going to a richer language. That is the issue to which I now turn.

17. General Hierarchies. What happens when we go from iterations D^α of D definable in a single classical fragment of L to iterations definable in the full L? A crucial point will be that the question of which syntactic operators on sentences count as iterations of D becomes "fuzzy": or put more precisely, we cannot in general assume

(O is an iteration of D) $\vee$ $\neg$(O is an iteration of D).

To see why this is so, let's define 'α^{th} iteration of D' as best we can for α larger than those discussed in the previous section.

To this end, we generalize the notion of a path. Let an *L-path* be some member p of P that assigns to each limit λ less than σ_p some L-definition of it. We know from Section 4 that the concept of an L-definition is "fuzzy", i.e. we can't in general assume excluded middle for claims of form 'u is an L-definition of v'; so there is no evident reason to assume it for formulas of form 'p is an L-path'. (Any L-path p is an ordinary function on $Pred_{\lim}(\sigma)$ for some unique σ, and we can reason about such functions in normal ways; but the question of which such functions count as L-paths is "fuzzy".) We can also generalize the Hierarchy-Existence Theorem:

> **General Theorem on Existence of Hierarchies:** There is an L_0-definable[53] function $\Phi : P \longrightarrow J$ such that
>
> p is an L-path $\vDash \Phi(p)$ (aka D_p) is a hierarchy of length σ_p.

(Again, the proof is deferred to the Appendix.)

But note that this theorem is a much weaker result than we had for the more restricted sorts of paths discussed in the previous section: it *doesn't* say that (for every p) *if* p is an L-path *then* D_p is a hierarchy extending up to σ_p. The function Φ "constructs" *something* from every $p \in P$, but because it may be "fuzzy" whether p is an L-path, it also may be "fuzzy" whether what's been constructed is a hierarchy of iterations of D. We can't even say that *if* p *is an* L*-path then* what we've constructed is such a hierarchy; all we can say is that if we're in a position to assert that p is an L-path then we're in a position to assert that what we've constructed is such a hierarchy.

We do have:

> **Corollary 1:** $\exists p$(p is an L-path of length σ) $\vDash \exists H$(H is a path-dependent hierarchy of iterations of D with length σ).

Proof: The previous theorem gives

p is an L-path of length $\sigma \vDash D_p$ is a hierarchy of D-iterations extending up to σ;

[53] The fact that Φ is definable in a classical fragment of L is of little intrinsic interest, but is essential to the proof: the proof employs an inductive definition that relies on the Replacement Schema, which is suspect once we leave demonstrably classical fragments of L.

existentially generalize over D_p in the conclusion, then use $\exists$-introduction on p. ■

But as with the previous theorem, we can't infer that *if* all predecessors of σ are L-definable *then* there is a hierarchy of D-iterations extending up to σ.

In addition, there is less of a connection than we might expect between the premise of the corollary and the claim that all predecessors of σ are L-definable. Certainly if there is an L-path of length σ, then all predecessors of σ are L-definable; but for the converse we are restricted to

> **Lemma:** All predecessors of σ are L-definable $\models$ There is an L-path of length σ.

Proof: The premise implies that $(\forall\lambda \in Pred_{\lim}(\sigma))(\exists y)(y$ is an L-definition of $\lambda)$; so by the "choice principle" mentioned at the end of Section 5 (and valid even for predicates not assumed classical), there is a function p with domain $Pred_{\lim}(\sigma)$ such that $(\forall\lambda \in Pred_{\lim}(\sigma))(p(\lambda)$ is an L-definition of $\lambda)$. ■

Call an ordinal **almost hereditarily L-definable** if all its predecessors are definable in L. (Note that ordinals that are (fully) hereditarily definable count as *almost* hereditarily definable as well.) Then

> **Corollary 2:**
>
> (Negative Part) If σ is not almost hereditarily L-definable then there are no path-dependent hierarchies of iterations of D with length σ.
>
> (Positive Part) σ is almost hereditarily L-definable $\models$ there are path-dependent hierarchies of iterations of D with length σ.

Proof: The Positive Part comes from the Lemma and Corollary 1, and the negative part is immediate. ■

We can also define a path-independent "hierarchy" ${D_{[L]}}^{\alpha}$ in complete analogy to how we defined the various ${D_{[L_R]}}^{\alpha}$; we'll see, though, that it is much less tractable. (Again I restrict the definition to 'v'-formulas, for simplicity.)

> **Definition of General Path-Independent "Hierarchy":** If A is any 'v'-formula, let ${D_{[L]}}^{\alpha}A$ be the formula (whose free variables are 'α' and 'v')
>
> $\exists p(p$ is an L-path $\wedge\ \alpha < \sigma_p \wedge$ the result of applying $[\Phi(p)](\alpha)$ to $\langle A\rangle$ is true of $v)$.

This defines the "hierarchy" for arbitrary α, but as with the hierarchies ${D_{[L_R]}}^{\alpha}$ of the previous section, there comes a point when it becomes ill-behaved: indeed, it eventually becomes trivial, in that for every sufficiently large α, ${D_{[L]}}^{\alpha}A$ is false of everything, for every formula A. What makes the situation much worse in this case is that we can say very little about where the breakdown occurs; indeed, this will turn out to be a "fuzzy" question.

We do have the following:

General Path-Independent Hierarchy Theorem:

Negative Part: If λ is a limit ordinal that is not definable in L, then for any $\alpha \geq \lambda$, ${D_{[L]}}^{\alpha}$ is trivial. Consequently, if σ is not almost hereditarily L-definable, then $D_{[L]} \restriction Pred(\sigma)$ fails (very badly!) to be a genuine hierarchy of iterations of D (or to be a genuine hierarchy of reasonable candidates for determinacy operators).

Positive Part: σ is almost hereditarily L-definable $\vDash D_{[L]} \restriction Pred(\sigma)$ is a genuine hierarchy of iterations of D.

(The proof of both parts is almost immediate from Corollary 2.) Note that since there are countable ordinals with undefinable predecessors, the negative part of this theorem implies that $D_{[L]}$ becomes very badly behaved *in the countable ordinals*. And the positive part, being in rule form rather than conditional form, is not enough to allow us to conclude that $D_{[L]}$ satisfies (RCL) up to any ordinal all predecessors of which are L-definable. For in order to universally generalize, you need a conditional to universally generalize on, and the theorem above does not license the strengthening to conditional form. The best we have is this: **for each limit σ that *we are in a position to assume* has only L-definable predecessors, we can take $D_{[L]} \restriction Pred(\sigma)$ to be a genuine hierarchy of iterations of D.** Once you *prove* that a given limit σ has only L-definable predecessors, the above results *converts the proof* into a demonstration that $D_{[L]} \restriction Pred(\sigma)$ is an adequate hierarchy.

18. Maximal Hierarchies? A question of great interest to revenge worries is whether there is a *maximal* hierarchy of iterations of D, that is, a σ for which $D_{[L]} \restriction Pred(\sigma)$ is adequate as a hierarchy (i.e. satisfies (RCL)) but $D_{[L]} \restriction Pred(\sigma + \omega)$ isn't. I will give a proof of the following "negative" answer:

Anti-Maximality Theorem: The assumption of such a maximal hierarchy of iterations of D is inconsistent.

In the course of this I will also establish the less interesting claim

Lemma for Anti-Maximality Theorem: The assumption of a maximal *L-definable* hierarchy of iterations of D is inconsistent.

I take the Lemma to have little intrinsic interest to the revenge problem: the proponent of revenge is sure to argue that the concepts that give rise to revenge problems aren't definable in the language. But the Anti-Maximality Theorem goes against not merely the assumption of a maximal *L-definable* hierarchy of iterations of D: the definability of the hierarchy doesn't enter into the result.

Proof of Lemma: If $D_{[L]} \restriction Pred(\sigma)$ is a genuine hierarchy of iterations of D then all predecessors of σ are definable in L (by negative part of Path-Independent Hierarchy Theorem). Assume that the hierarchy $D_{[L]} \restriction Pred(\sigma)$ is definable. Then σ is definable, as the length of the hierarchy. But then $\sigma + n$ is definable too for all finite n, and so every ordinal less than $\sigma + \omega$

is L-definable. But then the positive part of the Path-Independent Hierarchy Theorem tells us that $D_{[L]} \restriction Pred(\sigma+\omega)$ is adequate as a hierarchy of iterations of D. And this hierarchy is definable in L: for $\sigma+\ \omega$ is definable since σ is, and we've defined $D_{[L]}$. So $D_{[L]} \restriction Pred(\sigma)$ isn't maximal among the definable hierarchies of D-iterations. And so the assumption that it is maximal among the definable such hierarchies, and the existential generalization of that assumption, are inconsistent. ■

Proof of Theorem: Assume that $D_{[L]} \restriction Pred(\sigma)$ is a maximal hierarchy of iterations of D. Then we can define σ as the largest limit ordinal λ for which $D_{[L]} \restriction Pred(\lambda)$ is a hierarchy of iterations of D. But then we can define $D_{[L]} \restriction Pred(\sigma)$, so it is a maximal definable hierarchy, which is inconsistent by the Lemma. ■

It is important to be clear that from the inconsistency of the claim that there is a maximal hierarchy of iterations of D, it doesn't follow that there is no such hierarchy. (Any more than it follows from the inconsistency of the truth of the Liar sentence that the Liar sentence isn't true.) And indeed, the negation of the maximality claim is inconsistent too.[54] So the maximality claim has a status very much like that of the Liar sentence, in that the assumption that there *either is or isn't* a maximal hierarchy of iterations is inconsistent. We are in the realm of the "inherently fuzzy". (Anyone tempted to think that there *must* be a way to "unify" the hierarchies into a maximal one should re-read the response in Section 16 to the argument that there must be a way to "unify" the effectively classical hierarchies.)

19. Hyper-determinacy and Revenge. In Section 13 I distinguished three strengths of revenge worry. The weak form, which would not be totally devastating if substantiated, was that once one had a hierarchy of determinacy operators in a language L, one would be naturally led to a new "hyper-determinacy" operator, not definable in L but intuitively meaningful; since it is in an expansion L^* of L, the consistency proof for L wouldn't directly apply (though it might be extended to L^* in a completely mechanical way). The more serious revenge worries (the "intermediate strength" worry that a consistency proof for L^* would have to be quite different from that of L, and the "strong" worry that consistency for L^* could only be achieved by giving up the truth or satisfaction schema) seemed to depend on the view that such an operator would be idempotent (and even given that, the strong worry had little pre-theoretic support). We are now in a position to see that even the weak form of the worry was unfounded: there is no way to generate an understanding of a notion of hyper-determinacy from the hierarchy of determinacy notions that we have in the language.

We've seen that we can define within L a "hierarchy" $D_{[L]}$ that can be extended as far as one likes; but this is not a hierarchy of iterations of D—nor a

[54] Since each hierarchy has only L-definals ordinals in its domain, the lack of a maximal hierarchy would imply that there are arbitrarily large initial segments of the ordinals all members of which are definable in L, and that is absurd on cardinality grounds.

hierarchy of operators that are in any intuitive sense determinacy operators—since it eventually starts mapping every sentence into a falsehood. (This breakdown occurs somewhere in the countable ordinals.) Moreover, it is inconsistent to assume that there is a maximal initial segment of the ordinals on which $D_{[L]}$ behaves adequately—that is, a maximal fragment $D_{[L]} \restriction Pred(\sigma)$ such that for every limit ordinal $\lambda < \sigma$, ${D_{[L]}}^{\lambda}$ is equivalent to the "conjunction" of all the ${D_{[L]}}^{\beta}$ for $\beta < \lambda$. Since this is inconsistent, it seems that the best we can do if we want to avoid the danger of choosing a "hierarchy" that is inadequate is to choose a hierarchy that we can show to be adequate. But this will always be less than maximal. Given such a less than maximal hierarchy $\{{D_{[L]}}^{\alpha} \mid \alpha < \sigma\}$, we can always quantify over the operators in its range: calling a sentence A "hyperdeterminately true" would then be saying

(H_σ) For all α less than σ, $\langle {D_{[L]}}^{\alpha} A\rangle$ is true.[55]

(σ is bound to be definable in L if the fragment is adequate.) But in formulating (H_σ) we are in effect just going another step in a longer hierarchy *that is already in the language*. The general consistency proof for the theory of truth in L applies to everything expressible in the language, including iterations of D longer than those in the specific non-maximal hierarchy $\{{D_{[L]}}^{\alpha} \mid \alpha < \sigma\}$. So it certainly applies to sentences containing "hyper-determinacy" claims if that simply means claims of form (H_σ), for those are in the language

There is however another possibility to consider: to put it picturesquely, we can introduce the idea of a "fuzzy initial segment" of the full "hierarchy" $D_{[L]}$; in particular, the "initial segment consisting of all and only those ordinals that are in some adequate hierarchy", where again an adequate hierarchy (what I earlier called a genuine hierarchy) is one that obeys (RCL). This picturesque way of speaking makes no literal sense: since it is inconsistent to assume excluded middle for 'adequate', talk of an "initial segment" defined via the notion of adequacy is simply ill-defined. Even so, there is an idea behind the picturesque talk that can be made intelligible: that we define a "hyper-determinacy" operator using a quantifier restricted by the predicate 'is an adequate hierarchy'. I distinguish two versions of this:

$[H_{\supset}]A$ $\forall\alpha(\alpha$ is in the domain of an adequate hierarchy $\supset \langle {D_{[L]}}^{\alpha} A\rangle$ is true);

$[H_{\rightarrow}]A$ $\forall\alpha(\alpha$ is in the domain of an adequate hierarchy $\rightarrow \langle {D_{[L]}}^{\alpha} A\rangle$ is true).

It should be clear from the start that explaining hyper-determinateness in either of these ways can't possibly serve the purposes of the person who wants to argue for a revenge problem, even a weak one: for both $H_{\supset}$ and $H_{\rightarrow}$ are already in the language L. Since they are in the language, they can't possibly lead to paradox, given the general consistency proof. But it is worth seeing how

[55] This should really be written as "For all α less than σ, $\langle {D_{[L]}}^{v} A\rangle$ is true of α", but I trust the more readable notation in the text will not confuse.

the arguments for paradox fail. In the case of both operators, the failure of the paradoxical arguments points up the fact that "fuzzily restricted quantifiers" have to be treated with extreme care.

We saw in Section 10 that no operator E in the language can jointly satisfy four conditions. Those conditions were expressed in terms of models as (b_w), (c_w), (d) and idempotence; the corresponding inferential conditions are

(c_w) $\models EA \rightarrow A$

(d) $A \rightarrow \neg A \models \neg EA$

(b_w) $A, \neg EA \models \bot$

(Idem) $\models EA \rightarrow EEA$ (or equivalently, $\models \neg EEA \rightarrow \neg EA$)

The reason the conditions aren't jointly satisfiable (to transcribe an argument from Section 10 into inferential terms) is that for any such operator E we can formulate a sentence Q_E that asserts its own lack of E-truth, so that $\models Q_E \leftrightarrow \neg EQ_E$. Then since (d) implies $EQ_E \rightarrow \neg EQ_E \models \neg EEQ_E$ we'd have $EQ_E \rightarrow Q_E \models \neg EEQ_E$; so using ($c_w$), $\models \neg EEQ_E$. Idempotence would then yield $\models \neg EQ_E$, hence $\models Q_E$. (b_w) would then yield $\models \bot$, which is impossible. Given this general result, we know that neither $H_{\supset}$ nor $H_{\rightarrow}$ can possibly satisfy all of these four conditions. The question is, which of these do they satisfy, and which of the desirable additional conditions (a), (b) and (c) do they satisfy? (The D^{α} operators in adequate hierarchies satisfy the full (b) and (c) in addition to (a) and (d). I take that to be highly desirable: (b_w) and (c_w) were singled out only as minimal conditions for inconsistency with (Idem) and (d); and while the retreat from (c) to (c_w) is perhaps within the bounds of acceptability, the retreat from (b) to (b_w) would be a major one.)[56] I won't investigate (c), but will say enough about the other principles to show that neither $H_{\supset}$ nor $H_{\rightarrow}$ are operators that have much appeal.

(In discussing these matters I will occasionally state things in terms of standard set-theoretic models for L. It's worth noting that in the definitions of $H_{\supset}$ and $H_{\rightarrow}$ we could take all quantification over ordinals to be restricted to countable ordinals; for this reason, the points raised in Sections 8-9 about the "misleadingness of models" in dealing with sentences with unrestricted quantifiers will not affect the ability to infer truth from having value 1 and falsity from having value 0; conversely we can take *clear* truths to have value 1 in these models and *clear* falsehoods to have value 0.)

Let's start with $H_{\supset}$. Here the principles (c_w) and (d) are valid (in the extended sense introduced at the end of Section 5). The reason is that the

[56] Incidentally, we might want to add a further condition: that the operator E is to strengthen a given determinacy operator D, i.e.

(c_w*) $\models EA \rightarrow DA$.

Given this, E is bound to satisfy both (c_w) and (d), since D does. And we can now show that (c_w*) and (b_w) are together incompatible not only with (Idem) but with the weakened form of it

(W-Idem) $\models EA \rightarrow DEA$.

The proof is an obvious modification of the one given: (c_w*) yields $\models \neg DEQ_E$, which with (W-Idem) yields $\models Q_E$, which with (b_w) yields absurdity.

L-definability of 1 is valid, as is the equivalence of $True(\langle D^1_{[L]}A\rangle)$ with DA; so the conditional $H_\supset A \rightarrow DA$ is valid. Since D satisfies (c_w) and (d), it is then evident that $H_\supset$ does as well. (It can also be shown that $H_\supset$ must satisfy condition (a).)

An inspection of the above proof shows that because $H_\supset$ satisfies the principles (c_w) and (d), the sentence $H_\supset H_\supset Q_\supset$ must be false (where $Q_\supset$ is the "Liar sentence" corresponding to $H_\supset$). (Model-theoretically, $H_\supset H_\supset Q_\supset$ must have value 0 in any standard set-theoretic model.) $H_\supset Q_\supset$, on the other hand, will not have value 0 in any such model: for it is equivalent to $\neg Q_\supset$, and $Q_\supset$ won't have value 1. The reason: $Q_\supset$ says

$$\neg\forall\alpha(\alpha \text{ is in the domain of an adequate hierarchy} \supset \langle D_{[L]}{}^\alpha Q_\supset\rangle \text{ is true});$$

that is,

$$\exists\alpha(\alpha \text{ is in the domain of an adequate hierarchy} \wedge \neg(\langle D_{[L]}{}^\alpha Q_\supset\rangle \text{ is true})),$$

and the only way for this to have value 1 in a model that satisfies the truth schema is for it to be the case that for some α, $|\alpha$ is in the domain of an adequate hierarchy$| = 1$ and $\left|D_{[L]}{}^\alpha Q_\supset\right| = 0$. But $|\alpha$ is in the domain of an adequate hierarchy$| = 1$ only if α is in the domain of a clearly adequate hierarchy, in which case we can't have $|Q_\supset| = 1$ and $\left|D_{[L]}{}^\alpha Q_\supset\right| = 0$. The contradiction shows that $H_\supset Q_\supset$ can't be 0; so idempotence fails and paradox has been blocked. (By the reasoning of note 56, even "Weak-Idempotence" fails: that is, $DH_\supset$ is strictly stronger than $H_\supset$.)

So if the operators $D_{[L]}{}^\alpha$ are deemed problematic because not idempotent, $H_\supset$ offers no advantage. But in fact $H_\supset$ is far worse than the operators $D_{[L]}{}^\alpha$ in an adequate hierarchy, for it violates condition (b) in an extreme way: **For any sentence A whatever, $H_\supset A$ is "at best fuzzy".** (In any reasonable model for the language, no sentence of form $H_\supset A$ gets value 1.)[57] The reason is that $H_\supset A$ amounts to the claim "$\forall\alpha(\neg(\alpha$ is L-definable$) \vee D_{[L]}{}^\alpha$ is true)"; but this claim can't be assumed true even when A is a clear truth like '$0 = 0$', for when "it is fuzzy whether α is L-definable" it will be "fuzzy whether $D_{[L]}{}^\alpha$ is true", so that the disjunction "$\neg(\alpha$ is L-definable$) \vee D_{[L]}{}^\alpha A$ is true" will itself be "fuzzy". In short, $H_\supset A$ will never be clearly true.

We see, then, that though $H_\supset$ is a well-defined operator, it is quite a worthless one; it does not correspond to a notion of determinacy in any reasonable sense.

What about $H_\rightarrow$? Here the situation seems to be even worse. I say 'seems to be' because sentences of form $H_\rightarrow A$ are extremely complicated—this becomes evident when one consults the Appendix to see how the path-dependent hierarchies used in defining $D_{[L]}$ are themselves defined—and I have not been able to

[57] Of course, given that $H_\supset$ (i) is already in a language that has a G-semantics and (ii) is not syntactically restricted in its application, it must be value-functional; and this means that if there is any sentence A with value 1 for which $|H_\supset A| < 1$, this *must* be so for every sentence with value 1.

come up with a rigorous argument settling exactly how $H_{\rightarrow}$ behaves. However, I think there is *very* strong reason to believe the following:

1. In many G-logics, including all the published ones and all possible ones that satisfy the very natural law (III_s) of Section 10, $H_{\rightarrow}$ is a completely trivial operator, in that $H_{\rightarrow}A$ is clearly false for every sentence A. (So in a standard set-theoretic model of such a logic, $|H_{\rightarrow}A|$ is always 0, even if A has value 1.)

2. In any other G-logic, $H_{\rightarrow}$ will share the problem of $H_{\supset}$: $|H_{\rightarrow}A|$ will never have value 1, creating a serious disadvantage as compared with the genuine hierarchies of form $D^{\alpha}_{[L]}$. In these logics, $H_{\rightarrow}$ will also fail to be idempotent, thus destroying the entire rationale for going beyond hierarchies of form $D^{\alpha}_{[L]}$. ($H_{\rightarrow}$ as defined may also fail to satisfy (c_w), though that problem could be fixed by conjoining the above definiendum for $H_{\rightarrow}A$ with A.)

To substantiate 2, it would suffice to show that in any standard set-theoretic model for L, there are ordinals α for which $|\alpha$ is in the domain of an adequate hierarchy$| \not\leqslant |\exists p(p$ is a path of length greater than $\alpha \wedge \langle D^{\alpha}_{p}A\rangle$ is true)|;[58] to substantiate 1, it would suffice to show that in the semantics for the logics mentioned in 1 we can strengthen this, replacing '$\not\leqslant$' by '$>$'. And I think a strong case can be made for these claims, though I will not attempt it here since the case depends on a careful look at the way in which the $D^{\alpha}_{p}A$ are defined in the Appendix. (The fact underlying the plausibility of the claims is that $|D^{\alpha}_{p}A|$ is evaluated by looking at $|D^{\beta}_{p}A|$ for all β that precede some ordinal μ for which $|\mu$ satisfies $\_\_\_| > 0$, where the blank is filled by the "attempted definition of α" that p assigns to α. For ordinals α for which $|\alpha$ is in the domain of an adequate hierarchy| is between 0 and 1 and hence the "attempted definition" is not clearly adequate, we can expect this to include ordinals β much bigger than α, perhaps some for which $|\exists p(p$ is a path of length greater than $\beta)|$ is 0 and hence $|D^{\beta}_{p}A|$ will be 0.)

I've claimed only a *strong case* for the claims 1 and 2. But a strong case isn't a proof; what if I'm wrong? If I'm wrong, that would have some interesting ramifications: it would mean that the range of iterations of D available in the language is far richer than Section 17 might have suggested. More particularly, we could continue the sequence of non-idempotent operators even further than was done in the general hierarchies considered there. This would not however change anything substantive in what I've said: by the results of Section 11, the new hierarchy would never lead to idempotence as long as we could iterate in accordance with (RCL); and in accordance with the result of Section 18, there would be no maximal hierarchy of iterations of $H_{\rightarrow}$. And so introducing the operator $H_{\rightarrow}$ wouldn't have served the purpose that an advocate of "revenge" or of a "unified determinacy operator" intended. (It might motivate such a person to introduce a new *hyper-hyper* determinacy operator; but obviously nothing of philosophical significance could be gained by further travel down this road.)

[58] In analogu with note 55, I use $|F(\alpha)|$ as a more readable notation for $|F(v)|_{\alpha}$.

The crucial point is that even if my conjecture that $H_{\rightarrow}$ fails to meet the conditions for being a determinacy operator, it couldn't possibly constitute even a weak revenge problem, because $H_{\rightarrow}$ is already in the language. $H_{\rightarrow}$ avoids the most obvious threat of paradox by failing to be idempotent. And the general consistency result shows that it can't possibly lead to any paradox, because it is in L.

20. Conclusion. The discussion of the last few sections gives strong reason to think that we simply have no conception of any genuine hyper-determinacy operator that isn't definable in L: the closest we can come is operators like $H_{\supset}$ and $H_{\rightarrow}$ that *are* definable in L but that don't behave like a hyper-determinacy operator (or any sort of determinacy operator) ought to behave.

I haven't tried to argue that there is no intelligible expansion of our understanding of a hierarchy of determinacy operators. Indeed, it is clear from my formal constructions that if we were to expand our mathematical language in such a way that countable ordinals not hereditarily definable in our current L (or even, not *clearly* hereditarily definable) were to become clearly hereditarily definable, then that would expand our conception of the hierarchy: it would enable us to make stronger "iterated determinacy" claims than we can make today. But such an expansion of mathematics couldn't be simply a matter of defining new concepts in terms of current vocabulary, it would have to involve coming to have concepts that can't be clearly defined in the existing language; and achieving such an expansion is no simple task.[59]

In particular, part of the upshot of my argument is that such an expansion can't be achieved simply by reflecting on the hierarchies of determinacy operators already definable in the language. The thought that reflecting on such hierarchies leads to a concept of hyper-determinacy that transcends the language is simply an illusion, an illusion created by the failure to appreciate that if we try to quantify over all the determinacy operators definable in the language, the range of the quantification is indeterminate.

This, I think, defeats the (slightly vague) weak revenge worry of Section 13. *A fortiori* (and much more important), it defeats the idea that reflection on such hierarchies leads to a concept of an *idempotent* hyper-determinacy operator of the sort that might support an intermediate-strength (or perhaps even strong) revenge worry.

Of course, one might think that there is a case for thinking that we can understand an idempotent determinacy argument that is independent of reflecting on the hierarchies we can define. I've mentioned four possible grounds for this thought, three in Section 13 and another in Section 12, and I'll now add a fifth.

- One is the model-theoretic revenge argument, which I believe I have refuted in section 9.

[59] One might wonder about imaginary beings with an *uncountable* language that contained a name for every countable ordinal. The results in this paper could be extended to them: it's simply that the hierarchies would extend into the uncountable ordinals.

- Another is the thought that excluded middle holds generally. This is certainly not a view I can claim to have refuted in this paper: all I have tried to do (in section 2) is to sketch the costs that the semantic paradoxes raise for such a view, and to elaborate a view that seems on balance to have less drastic costs. (Of course, one may evaluate costs differently: *de gustibus non disputandum.*)

- A third one is that excluded middle should at least hold for claims of the form 'O is a reasonable candidate for being a determinately operator'. If that were true, we should be able to unproblematically quantify over all reasonable candidates for determinately operators, to produce a hyper-determinately operator that obeys excluded middle; it is then a short step to idempotence. But why think excluded middle holds for claims of this sort? The discussion in the last few sections provides strong reasons to doubt this supposition, and its hard to imagine a case for the supposition that doesn't rely either on excluded middle generally or on the thought that the only reasonable candidate for a determinacy operator could be read off the model theory.

- A fourth argument (the one not directly mentioned before) is that even if excluded middle doesn't hold for claims of the form 'O is a reasonable candidate for being a determinately operator', still we should be able to quantify over all reasonable candidates for determinately operators to produce a hyper-determinately operator; it may not obey excluded middle, but it might be thought to be idempotent on other grounds. But from the results of the preceding section, I think we can reasonably extrapolate to the view that there is little reason to expect such an operator to be idempotent, and little reason even to think that it will obey minimal conditions for being a determinacy operator.

The final argument (the one from Section 12) is perhaps the one with most intuitive force: it is that we *just need* a unified notion of determinacy or defectiveness. Note however that this argument cannot very well be advocated by the classical theorist, since the classical theorist has no such unified notion either. Nor can it very well be advocated by the proponent of any other solution to the paradoxes in which such a notion is unavailable. Indeed, I'm not sure that there are any demonstratively consistent theories (or even non-trivial dialetheic ones) that have such a notion available and hence are in a position to advocate this argument. I'm willing to concede (for the moment anyway) that it would be a point in favor of a solution to the paradoxes that it had a unified notion of defectiveness. If there are ways to achieve this that don't have overwhelming costs, they should be developed and weighed against the solutions to the paradoxes sketched here.[60]

[60] A number of people have tried to persuade me over the last few years that the "revenge-immune" account in [3] doesn't really evade revenge. I should especially mention Graham Priest, who has mostly pressed model-theoretic revenge (see [16]) and Kevin Scharp, who has

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

[3] Hartry Field. A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic*, 32:139–177, 2003.

[4] Hartry Field. The semantic paradoxes and the paradoxes of vagueness. In JC Beall, editor, *Liars and Heaps*. Oxford University Press, 2003.

[5] Hartry Field. The consistency of the naive theory of properties. *Philosophical Quarterly*, 54:78–104, 2004.

[6] Hartry Field. Maudlin's truth and paradox. *Philosophy and Phenomenological Research*, 2005.

[7] Hartry Field. Variations on a theme by yablo. In JC Beall and Brad Armour-Garb, editors, *Deflationism and Paradox*. Oxford University Press, 2005.

[8] Hartry Field. Compositional principles versus schematic reasoning. *The Monist*, 89, 2006.

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

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

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

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

[13] Tim Maudlin. *Truth and Paradox*. Oxford University Press, Oxford, 2004.

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

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

pressed the "incompatibility qualm" and the "counterintuitiveness qualm" of Section 12 (see Section A.5 of [19]). It was in thinking about what Scharp says that I was led to realize that the discussion of the hierarchy of determinacy operators in [3] was insufficiently inclusive, and needed to be extended into the "fuzzily definable" ordinals. I'm also very grateful to Josh Schechter for carefully reading an earlier draft and providing valuable comments that led me to substantial improvements.

[16] Graham Priest. Spiking the field artillery. In JC Beall and Brad Armour-Garb, editors, *Deflationism and Paradox*. Oxford University Press, 2005.

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

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

[19] Kevin Scharp. *Truth and Aletheic Paradox*. dissertation, University of Pittsburgh, 2005.

[20] Scott Soames. *Understanding Truth*. Oxford University Press, Oxford, 1999.

[21] Philip Welch. Ultimate truth v. stable truth. *Journal of Philosophical Logic*.

[22] Stephen Yablo. New grounds for naive truth theory. In JC Beall, editor, *Liars and Heaps*. Oxford University Press, 2003.

Appendix: Proof of the Hierarchy-Existence Theorems

It's simplest to directly prove the existence of hierarchies of operators on formulas with a single free variable 'v' ('v'-formulas). The Reasonability Conditions in the definition of the hierarchy are then modified in the obvious way: we restrict to operations on 'v'-formulas, and in (RCL) we speak of satisfaction by objects instead of by assignment functions. From such a "restrictive hierarchy" of operators on 'v'-formulas, a more general hierarchy of operators on all L-formulas could be obtained: for by a minor extension of the ideas of note 4 we could define in L_0 a function that takes operations on 'v'-formulas into corresponding operations on all L-formulas. (I spare you the details.)

The Hierarchy Existence Theorems of Sections 15 and 17 (modified in this way to apply to hierarchies of operators on 'v'-formulas) follow almost directly from a technical lemma. Take P to be as defined in Section 15: recall that it is an L_0-definable set each member p of which is a function with domain the set of limit ordinals that precede σ_p, for some limit ordinal σ_p that may depend on p.

> **Hierarchy-Construction Lemma:** There is an L_0-definable function W with domain $\{< p, \alpha >| \, p \in P \wedge \alpha < \sigma_p\}$ that satisfies the following conditions:

1. For any $p \in P$, $W(p, 0)$ is the identity operator on 'v'-formulas.
2. For any $p \in P$ and any $\alpha < \sigma_p$, $W(p, \alpha + 1)$ is $det(W(p, \alpha))$.

3. For any $p \in P$ and any limit ordinal $\lambda < \sigma_p$, $W(p, \lambda)$ is an operator on 'v'-formulas such that for any 'v'-formula x, $[W(p, \lambda)](x)$ is equivalent to the 'v'-formula $Z_{p,\lambda}(x)$ that results from substituting $p(\lambda)$ and the standard name of x into the blanks in

   $\forall\beta[\exists\mu(\_\_\_ \wedge \beta<\mu) \rightarrow$ the result of applying $W(p, \beta)$ to $\_\_\_$ is true of v.

So writing $W(p, \alpha)$ as $[\Phi(p)](\alpha)$, we immediately obtain that for every $p \in P$, $\Phi(p)$ satisfies the reasonability conditions for zero and successors (again, modified to apply to operators on 'v'-formulas). For the limit condition, on the other hand, we get the horrible-looking

> For any $p \in P$ and any limit ordinal $\lambda < \sigma_p$, $[\Phi(p)](\lambda)$ is an operator on 'v'-formulas such that for any 'v'-formula x, $[[\Phi(p)](\lambda)](x)$ is equivalent to the 'v'-formula $Z_{p,\lambda}(x)$ that results from substituting $p(\lambda)$ and the standard name of x into the blanks in
>
> $\forall\beta[\exists\mu(\_\_\_ \wedge \beta<\mu) \rightarrow$ the result of applying $[\Phi(p)](\beta)$ to $\_\_\_$ is true of v.

However, from the assumption that p is an L-path, what results from filling in the first blank is true of λ and nothing else, so $Z_{p,\lambda}(x)$ is satisfied by just those objects that satisfy all of the results of applying $[\Phi(p)](\beta)$ to x, for each $\beta < \lambda$. In other words, $\Phi(p)$ satisfies (RCL). Similarly for L_R-paths; but since we have excluded middle for 'L_R-path' (unlike for 'L-path), we can in the case of L_R convert the proof to the proof of the conditional: if p is an L_R-path then $\Phi(p)$ satisfies (RCL).

It remains only to prove the technical lemma.

Proof of Hierarchy-Existence Lemma: The obvious idea for proving the Lemma is to define the function W using transfinite recursion. But a *direct* recursive definition of W seems impossible, because of the fact that condition (3) of the Lemma doesn't just *use* W (as in a normal recursive definition) but *mentions* it (by referring to a formula that contains it). So instead, I will recursively define a more generalized function F (with little intuitive meaning, I regret to say), then use a fixed point argument to get the desired W.

Let Y be the set of formulas of L that have only the two variables 'β', and 'z' free. We want to recursively define a function $F(p, \alpha, e)$ for $p \in P$, $e \in Y$ and $\alpha < \sigma_p$, whose values are operators on 'v'-formulas. The idea is that if we then instantiate on an appropriate instance e_0, then the formula $F(p, \alpha, e_0)$ will serve as the desired W (and so for any specific L_0-path p, $F(p, \alpha, e_0)$ will serve as the desired hierarchy).

The recursive definition:

- For any $p \in P$ and $e \in Y$, let $F(p, 0, e)$ be the identity operator on 'v'-formulas.

- For any $p \in P$ and $e \in Y$ and any $\alpha<\sigma_p$, let $F(p, \alpha+1, e)$ be $det(F(p, \alpha, e))$.
- For any $p \in P$ and $e \in Y$ and any limit ordinal $\lambda<\sigma_p$, let $F(p, \lambda, e)$ be the operator that assigns to each 'v'-formula x the result of substituting $p(\lambda)$, e, and the standard name of x in that order into the blanks in the following schema:

  $\forall z \forall \beta[\exists \mu(\_\_\_ \wedge \beta<\mu) \wedge z$ is a syntactic operator on 'v'-formulas $\wedge$ $\_\_\_$ $\rightarrow$ the result of applying z to $\_\_\_$ is true of $v]$.

It should be noted that recursive definition is not unrestrictedly valid in L: it depends on the Replacement Schema, which is valid only in the context of excluded middle. But there is no problem with this particular recursive definition, for it is given in the 'true'-free fragment L_0. (The expression 'true' does occur here, but only in a sentence that is mentioned rather than used; it's mere syntax.) The recursive definition can be converted to an explicit definition of a relation $F(p, \alpha, e) = z$. Obviously for any particular e_0 that we restrict to, the first two bulleted conditions of the Lemma will be satisfied (by virtue of the corresponding conditions of the inductive definition); the task is to choose an e_0 that will make the third condition satisfied as well.

To this end, we now employ the Gödel-Tarski fixed point theorem on the formula '$F(p, \beta, e) = z$', to get a function $W(p, \beta)$ (defined in L_0) for which

$\forall p \forall \beta[W(p, \beta) = F(p, \beta, \langle W(p, \beta) = z \rangle)]$.[61]

Using '$W(p, \beta) = z$' to instantiating the e in above recursive definition, the limit condition of the definition yields

- For any $p \in P$ and any limit ordinal $\lambda<\sigma$, $W(p, \lambda)$ is an operator that assigns to each 'v'-formula x the result of substituting $p(\lambda)$, the definition of '$W(p, \beta) = z$', and the standard name of x in that order into the blanks in the following schema:

  $\forall z \forall \beta[\exists \mu(\_\_\_ \wedge \beta<\mu) \wedge z$ is a syntactic operator on L-formulas $\wedge$ $\_\_\_$ $\rightarrow$ the result of applying z to $\_\_\_$ is true of $v]$.

So for any p, $[W(p, \lambda)](x)$ is equivalent to the result of substituting $p(\lambda)$ and the standard name of x into the blank in

$\forall \beta[\exists \mu(\_\_\_ \wedge \beta<\mu) \rightarrow$ the result of applying $W_\sigma(p, \beta)$ to $\_\_\_$ is true of $v]$,

which is Condition (3).

[61] The most familiar form of the fixed point theorem applies to formulas. Applying it to the formula '$F(h, \beta, e) = z$', we get a three-place formula $G(h, \beta, z)$ of L_0 such that

$\forall z[G(h, \beta, z) \leftrightarrow F(h, \beta, < G(h, \beta, z) >) = z]$.

But $G(h, \beta, z)$ defines a function; writing it as $W(h, \beta) = z$, we get the claim in the text.