A Revenge-Immune Solution to the Semantic Paradoxes

Hartry Field*

August 29, 2002

Abstract

The paper offers a solution to the semantic paradoxes, one in which (1) we keep the unrestricted truth schema "True($\langle A \rangle$) $\leftrightarrow A$", and (2) the object language can include its own metalanguage. Because of the first feature, classical logic must be restricted, but full classical reasoning applies in "ordinary" contexts, including standard set theory. The more general logic that replaces classical logic includes a principle of substitutivity of equivalents, which with the truth schema leads to the general intersubstitutivity of True($\langle A \rangle$) with A within the language.

The logic is also shown to have the resources required to represent the way in which sentences (like the Liar sentence and the Curry sentence) that lead to paradox in classical logic are "defective". We can in fact define a hierarchy of "defectiveness" predicates within the language; contrary to claims that any solution to the paradoxes just breeds further paradoxes ("revenge problems") involving defectiveness predicates, there is a general consistency/conservativeness proof that shows that talk of truth and the various "levels of defectiveness" can all be made coherent together within a single object language.

1 Introduction

Kripke's theory of truth ([9]), in its strong Kleene version (where the law of excluded middle is not valid without restriction), has a very nice feature: True($\langle A \rangle$) (the assertion of a sentence A that it is true) is everywhere intersubstitutable with A, for any sentence A whatever (in a language without either indexicals, ambiguities, etc., or quotation marks, attitude contexts, etc.). But an unpleasant feature of the theory is that there is no biconditional $\leftrightarrow$ for which $A \leftrightarrow A$ is a logical law, or (equivalently, given the intersubstitutivity property) for which True $(\langle A \rangle) \leftrightarrow A$ holds in general. Indeed, there is no conditional $\rightarrow$ for which $A \rightarrow A$ is a logical law, or for which either True $(\langle A \rangle) \rightarrow A$ or its converse hold generally.

I will propose a natural extension of Kripke's theory to include a new conditional $\rightarrow$, satisfying $A \rightarrow A$, such that even in the extended language the intersubstitutivity of True $(\langle A \rangle)$ with A holds; so if $\leftrightarrow$ is defined from $\rightarrow$ in the

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

obvious way, we get both $A \leftrightarrow A$ and $\mathsf{True}(\langle A \rangle) \leftrightarrow A$ as general theorems. In short, we get *the naive theory of truth*: the truth schema *together with* intersubstitutivity. The new conditional obeys many of the laws we expect of a conditional, such as modus ponens and contraposition; some laws, such as the inference from $A \rightarrow (B \rightarrow C)$ to $A \wedge B \rightarrow C$, fail, but this is inevitable given the Curry paradox (mentioned below).

I gave a different, rather artificial, way of adding such a conditional to Kripke's theory in [5]; the one that follows leads to a different logic, stronger in important respects though also weaker in others. The one offered here has several advantages. First, it is based on a very natural semantics. Second, the new conditional is equivalent to the classical conditional in those contexts where excluded middle is assumed for the antecedent and consequent. Third and probably most important, the new conditional can be used to show that the theory is not subject to "revenge problems".

More fully, the addition of the new conditional operator to the language allows for the definition of a natural "determinately operator", so that we can consistently handle "extended paradoxes", such as sentences that assert of themselves that they are not determinately true. I will discuss the resolution of a number of such extended paradoxes—including some that notoriously make trouble for other attempted resolutions of the paradoxes—near the end of the paper.

As we'll see, we can in fact define a transfinite hierarchy of stronger and stronger determinately operators within the language. If we think of a determinately operator as attaching to a truth predicate to yield a predicate of "strong truth", we can think of the theory as providing an account of "stronger and stronger truth predicates". But unlike most approaches that allow a hierarchy of "truth predicates", no infinite hierarchy of metalanguages is required. Indeed there need be no distinction between metalanguage and object language at all: if the object language is rich enough to include standard set theory (ZFC) and a single notion of truth that obeys the truth schema (and of course the Kleene connectives and the new $\rightarrow$), then all these other "truth predicates" are definable within the object language.

2 The Construction

Let's start with a ground language $\mathcal{L}$ that is adequate to arithmetic, in the sense that one can explicitly define in $\mathcal{L}$ a predicate 'natural number' and predicates corresponding to the usual basic arithmetical notions ('is zero', 'is the successor of', 'is the sum of' and 'is the product of'). As usual, the point of wanting arithmetic is that it can be used to develop the formal syntax of $\mathcal{L}$ and certain extensions of it, by Gödel numbering. I also suppose that the language contains, or has definable in it, the usual vocabulary for talking about finite sequences (viz., the 1-place predicate 'x is a finite sequence', the 2-place predicates 'n is the length of x', and the 3-place predicate 'b is the n^{th} member of x'); when $\mathcal{L}$ is just the language of arithmetic this is no extra assumption, but in the general case that additional vocabulary is required as a basis for semantic talk, e.g. of satisfaction of formulas by finite sequences of objects.

Let $\mathcal{L}^+$ be the result of adding to $\mathcal{L}$ both a 1-place predicate 'True' and a 2-place operator $\rightarrow$ on formulas. 'Free variable' is defined in the usual way: in particular (in contrast to the preferred treatment of $\rightarrow$ in [5]), variables that are free in A or in B also count as free in $A \rightarrow B$.

Let M be a classical model for $\mathcal{L}$. (The stipulation that M be classical reflects the assumption that classical logic is appropriate to $\mathcal{L}$ if not to $\mathcal{L}^+$. It is arguable that one should relax this assumption when $\mathcal{L}$ contains vague terms, and in fact the account here easily generalizes to where M is a 3-valued model. But let's keep things simple.) I will assume that "the arithmetical part of M" is a standard model of arithmetic; where by "the arithmetical part of M" I mean the submodel whose domain is the set of objects that satisfy the definition of 'natural number'. So M must be infinite. Analogously, I assume that M validates the usual theory of finite sequences.[1] (Given the standardness of the arithmetical part of the model, anything satisfying 'finite sequence' must have a genuinely finite number as its length.) Otherwise, M is arbitrary.

It will simplify the development (or at least the notation) if we assume that for every object in M there is a name in $\mathcal{L}$ and hence in $\mathcal{L}^+$.[2] This is no real loss of generality: if the original $\mathcal{L}$ and $\mathcal{L}^+$ don't have that, we can pass to an $\mathcal{L}^\star$ and $\mathcal{L}^{+\star}$ that do. Assuming the original $\mathcal{L}$ had a finite or countable vocabulary, the new $\mathcal{L}$ and $\mathcal{L}^+$ will have a vocabulary of the same cardinality as M. I now suppose that the syntax of $\mathcal{L}^+$ (the new one, i.e. $\mathcal{L}^{+\star}$) is developed within $\mathcal{L}$, in a standard way. (Because $\mathcal{L}^+$ may be uncountable, we can't just use Gödel numbering properly so called; but we can give name-free formulas ordinary Gödel numbers, and assign to sentences finite sequences whose first member is the Gödel number of a name-free formula and whose other members are names. I'll call these 'Gödel codes', and will sometimes identify sentences with their Gödel codes.)

Let Ω be the initial ordinal of the cardinality that immediately succeeds that of M. For any ordinals α and σ, with $\sigma \leq \Omega$, I will now extend M to a 3-valued model $M_{\alpha,\sigma}$ of $\mathcal{L}^+$. The three values I call 1, 0, and $\frac{1}{2}$. (Please do *not* think of these as meaning 'true', 'false', and 'neither true nor false'. A slightly better rendition would be 'determinately true', 'determinately false', and 'neither determinately true nor determinately false', though as we shall eventually see, this is inaccurate as well.) The models are to be constructed in the lexicographical order; that is, $\langle \alpha, \sigma \rangle \preceq \langle \alpha^\star, \sigma^\star \rangle$ iff either $\alpha < \alpha^\star$ or both $\alpha = \alpha^\star$ and $\sigma \leq \sigma^\star$. I will eventually argue that there is a non-zero Δ such that for all $\beta > 0$, all the $M_{\Delta \cdot \beta, \Omega}$ coincide. The resulting $M_{\Delta,\Omega}$ (which is unique, since any two ordinals have a common non-zero right multiple) I'll call $M^\star$, and it will be the desired extension of M to $\mathcal{L}^+$. We can think of the different values of $\alpha < \Delta$ as representing different "super-stages" toward the construction of the "final answer"

[1] I.e., that every finite sequence has exactly one length, which is a positive integer; that for every sequence x and every positive integer k less than or equal to its length, there is exactly one k^{th} member of x; and for each natural number n the claim

> $\forall x_1 ... \forall x_n \exists! y$[$y$ is a finite sequence $\wedge$ n is the length of $y \wedge x_1$ is the 1^{st} member of $y \wedge ... \wedge x_n$ is the n^{th} member of y].

[2] The sole point of this is to avoid having to assign semantic values to formulas with free variables; this in turn avoids having to relativize semantic value to an assignment of members of the domain of M to the free variables, which is notationally messy.

$M^\star$ as to what the 3-valued model should be; the different values of σ for a fixed α represent different "mini-stages" within a super-stage, and I will often drop the ordinal for the mini-stages out of the notation.

Notation: $[M]$ is the domain of M. SENT is the subset of $[M]$ consisting of the Gödel codes of sentences of $\mathcal{L}^+$; neg is the operation on Gödel codes corresponding to negation. If t is a variable-free term of $\mathcal{L}^+$ (hence of $\mathcal{L}$), $den(t)$ is its denotation in M; if p is a predicate of $\mathcal{L}$, p_M is its extension in M. If A is a formula of $\mathcal{L}^+$, $A(x/t)$ is the result of substituting t for all free occurrences of x (changing the bound variables of A as necessary to avoid conflicts). $\langle A\rangle$ is the name of the Gödel code of A. $[\beta, \alpha)$ is the set of ordinals $\geq \beta$ and $< \alpha$.

Each new model $M_{\alpha,\sigma}$ has the same domain as M, and all the terms denote what they denote in M. $M_{\alpha,\sigma}$ assigns semantic values in $\{0, \frac{1}{2}, 1\}$ to sentences of $\mathcal{L}^+$ as follows:

1. If p is an n-place predicate *of* $\mathcal{L}$ and t_1,...,t_n are variable-free terms:
$$|p(t_1, ..., t_n)|_{\alpha,\sigma} = \begin{cases} 1 & \text{iff } \langle den(t_1), ..., den(t_n)\rangle \in p_M \\ 0 & \text{otherwise.} \end{cases}$$
2. If t is a variable-free term,
$$|\mathsf{True}(t)|_{\alpha,\sigma} = \begin{cases} 1 & \text{iff for some sentence } A, den(t) \text{ is the Gödel code of } A \\ & \quad \text{and } (\exists\delta < \sigma)(|A|_{\alpha,\delta} = 1); \\ 0 & \text{iff for some sentence} A, den(t) \text{ is the Gödel code of } A \\ & \quad \text{and } (\exists\delta < \sigma)(|A|_{\alpha,\delta} = 0), \text{ or there is no sentence } A \\ & \quad \text{for which } den(t) \text{ is the Gödel code of } A; \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$
3. $|\neg A|_{\alpha,\sigma} = 1 - |A|_{\alpha,\sigma}$
4. $|A \wedge B|_{\alpha,\sigma} = min\{|A|_{\alpha,\sigma}, |B|_{\alpha,\sigma}\}$
5. $|A \vee B|_{\alpha,\sigma} = max\{|A|_{\alpha,\sigma}, |B|_{\alpha,\sigma}\}$
6. $|\forall xA|_{\alpha,\sigma} = min\{|A(x/t)|_{\alpha,\sigma} \mid t \text{ is a variable-free term }\}$
7. $|\exists xA|_{\alpha,\sigma} = max\{|A(x/t)|_{\alpha,\sigma} \mid t \text{ is a variable-free term }\}$
8. $$|A \rightarrow B|_{\alpha,\sigma} = \begin{cases} 1 & \text{iff } (\exists\beta < \alpha)(\forall\gamma \in [\beta, \alpha))(|A|_{\gamma,\Omega} \leq |B|_{\gamma,\Omega}) \\ 0 & \text{iff } (\exists\beta < \alpha)(\forall\gamma \in [\beta, \alpha))(|A|_{\gamma,\Omega} > |B|_{\gamma,\Omega}) \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

This is a legitimate inductive definition (by complexity for fixed α, σ; in stages 2 and 8 we appeal to the values of formulas that may have greater complexity than the one whose value is at issue, but only for pairs $\langle\beta, \delta\rangle$ strictly prior to $\langle\alpha, \sigma\rangle$ in the $\preceq$ ordering). We need to look at its properties.[3]

[3]This construction is somewhat similar in spirit to that of Brady [2]; but Brady uses a different treatment of the conditional at each stage, which makes his construction monotonic, quite unlike what we have here. The conditional here, rather, is governed by a revision rule of the sort that Gupta and Belnap [6] use for the truth predicate. The results that emerge from the present construction are quite different from those that emerge from either Brady or Gupta and Belnap. (1) Gupta and Belnap get classical logic, and hence must restrict the truth schema. (2) Brady does not get the rule B3 below (or the Explosion rule that follows from it), and we will see that that rule plays a crucial role in shaping the character of the present theory.

Note that the value of $|A \rightarrow B|_{\alpha,\sigma}$ is always the same as $|A \rightarrow B|_{\alpha,0}$ (or $|A \rightarrow B|_{\alpha,\Omega}$); that is, it is completely independent of σ. This means that for each fixed α, the sequence of σ's is really just a standard Kripkean construction of a minimal fixed point over an assignment of the values $|A \rightarrow B|_{\alpha,0}$ to the conditionals. The first three results below simply make this more precise.

> **The Kripke monotonicity lemma:** For each α and each A, if $\sigma < \rho$ and $|A|_{\alpha,\sigma}$ has an integral value (0 or 1), then $|A|_{\alpha,\rho}$ has that same integral value.

Proof: Let $U_{\alpha,\sigma}$ be the set of sentences A such that $|A|_{\alpha,\sigma} = 1$. Then the set of sentences A such that $|A|_{\alpha,\sigma} = 0$ is $\{A \mid \neg A \text{ is in } U_{\alpha,\sigma}\}$; call this $U^{neg}_{\alpha,\sigma}$. We need that if $\sigma < \rho$ then $U_{\alpha,\sigma} \subseteq U_{\alpha,\rho}$ (from which it follows that $U^{neg}_{\alpha,\sigma} \subseteq U^{neg}_{\alpha,\rho}$). The proof is by induction on ρ (varying σ in the induction, but holding α fixed). The claim holds vacuously if ρ is 0. Suppose that $\rho > 0$. We show by a routine subinduction on the complexity of sentences A in the language that for all $\sigma < \rho$, if $A \in U_{\alpha,\sigma}$ then $A \in U_{\alpha,\rho}$ and if $A \in U^{neg}_{\alpha,\sigma}$ then $A \in U^{neg}_{\alpha,\rho}$. If A is a conditional, these are trivial since the value of any conditional is independent of the mini-stage. It's also trivial for atomic sentences of the ground language $\mathcal{L}$. If $\mathsf{True}(t) \in U_{\alpha,\sigma}$, then for some sentence B and some $\delta < \sigma$, $den(t)$ is the Gödel code of B and $B \in U_{\alpha,\delta}$; since $\delta < \rho$, $\mathsf{True}(t) \in U_{\alpha,\rho}$. If $\mathsf{True}(t) \in U^{neg}_{\alpha,\sigma}$, then either (i) for some sentence B and some $\delta < \sigma$, $den(t)$ is the Gödel code of B and $B \in U^{neg}_{\alpha,\delta}$; or (ii) $den(t)$ is not the Gödel code of a sentence. Either way, $\mathsf{True}(t)$ is in $U^{neg}_{\alpha,\rho}$. If $\neg B \in U_{\alpha,\sigma}$, $B \in U^{neg}_{\alpha,\sigma}$, so by induction hypothesis $B \in U^{neg}_{\alpha,\rho}$ and so $\neg B \in U_{\alpha,\rho}$; similarly if $\neg B \in U^{neg}_{\alpha,\sigma}$. Conjunctions, disjunctions and quantifications are easy. ■

> **Kripke Fixed Point Theorem:** For each α there is a $\sigma[\alpha] < \Omega$ such that for all $\sigma \geq \sigma[\alpha]$ and all A, $|A|_{\alpha,\sigma} = |A|_{\alpha,\sigma[\alpha]}$. (So in particular, $|A|_{\alpha,\Omega}$ is just $|A|_{\alpha,\sigma[\alpha]}$.)

Proof: The monotonicity lemma says that if $\sigma < \rho$ then $U_{\alpha,\sigma} \subseteq U_{\alpha,\rho}$ (where these are as in the proof of that lemma). Since the cardinality of the set of sentences is less than that of the predecessors of Ω, it can't be that the inclusion is proper for all ordinals less than Ω; so there is a $\sigma[\alpha] < \Omega$ with $U_{\alpha,\sigma[\alpha]+1} = U_{\alpha,\sigma[\alpha]}$. It follows that $U^{neg}_{\alpha,\sigma[\alpha]+1} = U^{neg}_{\alpha,\sigma[\alpha]}$. So $|A|_{\alpha,\sigma[\alpha]+1}$ has the same value as $|A|_{\alpha,\sigma[\alpha]}$; from which it follows that for all $\sigma \geq \sigma[\alpha]$, $|A|_{\alpha,\sigma} = |A|_{\alpha,\sigma[\alpha]}$. ■

> **Kripke Fixed Point Corollary:** For each α and each A, $|\mathsf{True}(\langle A \rangle)|_{\alpha,\sigma[\alpha]} = |A|_{\alpha,\sigma[\alpha]}$. (And $|\mathsf{True}(t)|_{\alpha,\sigma[\alpha]}$ is 0 when $den(t)$ is not a sentence.)

Proof: $|\mathsf{True}(\langle A \rangle)|_{\alpha,\sigma[\alpha]}$ is $|\mathsf{True}(\langle A \rangle)|_{\alpha,\sigma[\alpha]+1}$ by the fixed point theorem, which by Clause 2 above and monotonicity is 1 iff $|A|_{\alpha,\sigma[\alpha]}$ is 1 and 0 iff $|A|_{\alpha,\sigma[\alpha]}$ is 0, hence $\frac{1}{2}$ iff $|A|_{\alpha,\sigma[\alpha]}$ is $\frac{1}{2}$. ■

It easily follows that $\mathsf{True}(\langle A \rangle)$ is intersubstitutable with A when these are not within the scope of an $\rightarrow$, but in fact that restriction is unnecessary. More precisely:

> **Lemma on Substitutivity of Truth:** For any sentences A and B, if $B^\star$ results from B by replacing one or more occurrences of A by $\mathsf{True}(\langle A \rangle)$ then for any α, $|B|_{\alpha,\sigma[\alpha]} = |B^\star|_{\alpha,\sigma[\alpha]}$.

Proof: It obviously suffices to prove this for a single substitution, and to do that we use an induction on the depth of the embedding of the substituted occurrence of A in B. The basis (the case where B is A) is just the Fixed Point Corollary. The induction clauses corresponding to the connectives other than $\rightarrow$ hold by virtue of those connectives being degree-functional for any α and σ. The induction clause corresponding to $\rightarrow$ goes by a subinduction on α. ■

For *much* of what follows, there will be no need to make the mini-stages explicit, and so I will frequently drop them from the notation; that is, I'll let $|A|_\alpha$ abbreviate $|A|_{\alpha,\Omega}$, or equivalently, $|A|_{\alpha,\sigma[\alpha]}$. (The ordinal for mini-stages is rarely needed in calculation of semantic values: in determining the value of $|A|_{\alpha,\Omega}$, the only clause of 1-8 that requires a look at a mini-stage other than Ω is clause 2, and we can usually just use the Fixed Point Corollary for that. The proofs of some key theorems will require bringing in mini-stages, but we can do that as needed.)

Let's now look at the semantic values of the conditional. It is immediate from 8 that for any A and B, $|A \rightarrow B|_0$ is $\frac{1}{2}$, and that $|A \rightarrow B|_{\alpha+1}$ is either 1 or 0, depending on whether $|A|_\alpha \leq |B|_\alpha$. At limits λ, $|A \rightarrow B|_\lambda$ can have any of the three values; but we have the following continuity result:

> **Continuity Lemma for Conditionals:** At any limit λ, the value of a conditional is continuous: that is,
>
> $$|A \rightarrow B|_\lambda = \begin{cases} 1 & \text{iff } (\exists\alpha < \lambda)(\forall\beta \in [\alpha,\lambda))(|A \rightarrow B|_\beta = 1) \\ 0 & \text{iff } (\exists\alpha < \lambda)(\forall\beta \in [\alpha,\lambda))(|A \rightarrow B|_\beta = 0) \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

Proof: The Right to Left halves of the 1 and 0 clauses are straightforward: e.g., if $(\exists\alpha < \lambda)(\forall\beta \in [\alpha,\lambda))\,|A \rightarrow B|_\beta = 1$, then in particular $(\exists\alpha < \lambda)(\forall\beta \in [\alpha,\lambda))\,|A \rightarrow B|_{\beta+1} = 1$, so $(\exists\alpha < \lambda)(\forall\beta \in [\alpha,\lambda))\,|A|_\beta \leq |B|_\beta$, so $|A \rightarrow B|_\lambda = 1$. The Left to Right involves an induction: we suppose the result holds for all limits $< \lambda$ and we prove it for λ. Suppose $|A \rightarrow B|_\lambda = 1$; then for some $\alpha < \lambda$, $(\forall\beta \in [\alpha,\lambda))\,|A|_\beta \leq |B|_\beta$. Then $(\forall\beta \in [\alpha,\lambda))\,|A \rightarrow B|_{\beta+1} = 1$. So by induction hypothesis, $(\forall\beta \in [\alpha+1,\lambda))\,|A \rightarrow B|_\beta = 1$. The 0 case is analogous. ■

Define $||A||$ (the "ultimate value" of A) as

$$\begin{cases} 1 & \text{iff } (\exists\alpha)(\forall\beta \geq \alpha)(|A|_\beta = 1) \\ 0 & \text{iff } (\exists\alpha)(\forall\beta \geq \alpha)(|A|_\beta = 0) \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

It is clear that when either $||A||$ is 0 or $||B||$ is 1, then $||A \rightarrow B||$ is 1, and that when $||A||$ is 1 and $||B||$ is 0, $||A \rightarrow B||$ is 0. Also that if $||A \rightarrow B||$ is 1, then $||A|| \leq ||B||$. When $||A||$ is 1 and $||B||$ is $\frac{1}{2}$, or when $||A||$ is $\frac{1}{2}$ and $||B||$ is 0, the above leave open whether $||A \rightarrow B||$ is 0 or $\frac{1}{2}$; and both are possible.[4] On the $||A|| = ||B|| = \frac{1}{2}$ case, more in a moment.

[4]When $||A||$ is 1 and $||B||$ is $\frac{1}{2}$, $||A \rightarrow B||$ will be 0 iff there's an α after which $|B|_\alpha$ never takes value 1. The simplest case where this condition holds is the Liar sentence, which (as we'll see) has value $\frac{1}{2}$ at all stages; and a simple case where the condition fails is the Curry sentence, which as we'll see takes values 0 and 1 arbitrarily late. When $||A||$ is $\frac{1}{2}$ and $||B||$ is 0, $||A \rightarrow B||$ will be 0 iff there's an α after which $|A|_\alpha$ is never 0; again the Liar sentence is one case of this, and the Curry sentence isn't.

Note that as a consequence of the continuity lemma, we have that $||A \rightarrow B|| = 1$ iff for all sufficiently large α, $|A|_\alpha \leq |B|_\alpha$. (The right hand side directly gives only that $|A \rightarrow B|_\alpha = 1$ for all sufficiently large *successor* α; but the continuity lemma extends this to sufficiently large limit ordinals as well.) Similarly, $||A \rightarrow B|| = 0$ iff for all sufficiently large α, $|A|_\alpha > |B|_\alpha$.

A natural question is: are there ordinals α such that for every sentence A, $|A|_\alpha = ||A||$? I call such values of α *acceptable points*, and the nature of the theory depends very much on whether there are any, and on whether they occur arbitrarily late in the sequence of ordinals. For instance, without acceptable points, there would be no obvious reason to suppose that there couldn't be sentences A and B for which $||A \vee B|| = 1$ even though neither $||A|| = 1$ nor $||B|| = 1$: the disjunction has value 1 as long as there is a point past which $|B|_\alpha$ is 1 whenever $|A|_\alpha$ isn't 1, and it certainly isn't obvious that this requires that there is a point past which either $|A|_\alpha$ is always 1 or $|B|_\alpha$ is always 1. If on the other hand there are acceptable α, the Kleene rules hold at them, so this situation can't arise.

Another consequence of there being *arbitrarily big* acceptable points is that when $||A|| = ||B|| = \frac{1}{2}$, $||A \rightarrow B||$ is either $\frac{1}{2}$ or 1. (There are cases of both.) Reason: if $||A|| = ||B|| = \frac{1}{2}$, then if Δ is acceptable, $|A|_\Delta = |B|_\Delta = \frac{1}{2}$, so $|A \rightarrow B|_{\Delta+1}$ is 1; so if acceptable points occur arbitrarily late, $||A \rightarrow B||$ can't be 0. We see then (using also the remarks of three paragraphs back) that if there are arbitrarily big acceptable points, we have the following table of possible values for $\rightarrow$:

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

In the next section, I will give the simplest argument I have been able to find that there are indeed acceptable points, and that they occur arbitrarily late. (Indeed, I'll sketch a proof that there are acceptable points such that any non-zero right-multiple of them is acceptable.) A price of its simplicity is that the proof is less informative than one might like about the way the values of sentences change as the level increases toward an acceptable point. I believe a more informative proof should be possible, which would show among other things that acceptable fixed points (meeting the right-multiple condition) occur prior to Ω.[5] But the bare existence of arbitrarily late acceptable points is all that I require for my main claims, so I will limit myself to that.

3 Acceptable points

Let SENT be the set of sentences of $\mathcal{L}^+$. For any α, let $f(\alpha)$ be the function that assigns to each A in SENT the value $|A|_\alpha$ determined by the valuation rules. If $v = f(\alpha)$, I say that α *represents* v. And if $f(\alpha) = f(\beta)$ I say that α *is equivalent to* β.

[5] In Section 5 I will provide a detailed discussion of some examples; this should give a pretty good sense of how the values of sentences reach the values they take on at acceptable points, even in absence of the more informative proof.

Let FINAL be the set of functions v that are represented arbitrarily late, i.e. such that $(\forall\alpha)(\exists\beta \geq \alpha)(v = f(\beta))$.

Prop. 1: FINAL $\neq \emptyset$.

Proof: If it were empty, then for each function v from SENT to $\{0, \frac{1}{2}, 1\}$, there would be an α_v such that $(\forall\beta \geq \alpha_v)(v \neq f(\beta))$. Let θ be the supremum of all the α_v. Then for each function v from SENT to $\{0, \frac{1}{2}, 1\}$, $v \neq f(\theta)$. Since $f(\theta)$ itself is such a function, this is a contradiction. ■

Call an ordinal γ *ultimate* if it represents some v in FINAL; that is, if $(\forall\alpha)(\exists\beta \geq \alpha)(f(\gamma) = f(\beta))$.

Prop. 2: If α is ultimate and $\alpha \leq \beta$ then β is ultimate.

Proof: If $\alpha \leq \beta$, then for some δ, $\beta = \alpha + \delta$. Suppose α is ultimate. Then for any μ, there is an $\eta_\mu \geq \mu$ for which $f(\alpha) = f(\eta_\mu)$. But if $f(\alpha) = f(\eta_\mu)$, then $f(\beta) = f(\alpha + \delta) = f(\eta_\mu + \delta)$, and $\eta_\mu + \delta \geq \mu$; so β is ultimate. ■

Prop. 3: For any A, $||A|| = 1$ iff for every ultimate α, $|A|_\alpha = 1$; similarly for 0 and therefore for $\frac{1}{2}$.

Proof: Suppose that $||A|| = 1$, i.e. that there is an α such that $(\forall\beta \geq \alpha)(|A|_\beta = 1)$. For any ultimate ordinal γ, there is a $\beta \geq \alpha$ such that $f(\beta) = f(\gamma)$, hence in particular for which $|A|_\beta = |A|_\gamma$; so $|A|_\gamma = 1$. Conversely, suppose that $|A|_\alpha = 1$ for every ultimate α. Then by Prop. 2 we have $(\forall\beta \geq \alpha)(|A|_\beta = 1)$ whenever α is ultimate; since there are ultimate ordinals by Prop. 1, $||A|| = 1$. (The claim with 0 instead of 1 can be proved analogously, or deduced from the claim for 1 via negations.) ■

Given Prop. 3, the requirement that an ordinal α be acceptable amounts to the requirement that $\forall A[|A|_\alpha = 1$ iff $\forall\beta$ (if β is ultimate then $|A|_\beta = 1)]$ (and similarly for 0, though that's redundant). I now proceed to find an acceptable α.

Start with any ultimate ordinal τ, however large. Then every member of FINAL is represented by some ordinal $\geq \tau$; and since FINAL is a set rather than a proper class, and τ is ultimate, there must be a ρ such that $\tau + \rho$ is equivalent to τ and every member of FINAL is represented in the interval $[\tau, \tau+\rho)$. Finally, let Δ be $\tau + \rho \cdot \omega$.

Prop. 4: For any finite n and any $\alpha < \rho$, $f(\tau + \rho \cdot n + \alpha) = f(\tau + \alpha)$.

Proof: By choice of ρ, $f(\tau+\rho) = f(\tau)$. So for any β, $f(\tau+\rho+\beta) = f(\tau+\beta)$. If $\beta < \rho\cdot\omega$, write it as $\rho\cdot n+\alpha$, where $\alpha < \rho$. Then $f(\tau+\rho+\rho\cdot n+\alpha) = f(\tau+\rho\cdot n+\alpha)$; that is, $f(\tau + \rho \cdot (n + 1) + \alpha) = f(\tau + \rho \cdot n + \alpha)$. The result then follows, by induction on n. ■

Corollary to Prop. 4: For any n, every member of FINAL is represented in the interval $[\tau + \rho \cdot n, \tau + \rho \cdot (n + 1))$. ■

Prop. 5: For any sentences B and C, $|B \rightarrow C|_\Delta = ||B \rightarrow C||$.

Proof: Suppose $||B \rightarrow C|| = 1$. By Props. 3 and 2, $|B \rightarrow C|_{\alpha+1} = 1$ for all $\alpha+1$ in $[\tau, \Delta)$, so $|B|_\alpha \leq |C|_\alpha$ for all α in $[\tau, \Delta)$ (using the fact that Δ, i.e. $\tau+\rho\cdot\omega$,

is a limit); so $|B \rightarrow C|_{\Delta} = 1$. Similarly, if $||B \rightarrow C|| = 0$, then $|B \rightarrow C|_{\Delta} = 0$. It remains to prove the converses.

Suppose $|B \rightarrow C|_{\Delta} = 1$. Then for some $\alpha < \tau + \rho \cdot \omega$, we have that $(\forall \beta \in [\alpha, \tau + \rho \cdot \omega))(|B|_{\beta} \leq |C|_{\beta})$. Since $\alpha < \tau + \rho \cdot \omega$, there must be an n such that $\alpha < \tau + \rho \cdot n$. So $(\forall \beta \in [\tau + \rho \cdot n, \tau + \rho \cdot \omega))(|B|_{\beta} \leq |C|_{\beta})$. But by the corollary to Prop. 4, every member of FINAL is represented in $[\tau + \rho \cdot n, \tau + \rho \cdot \omega)$; so for every ultimate ordinal β, $|B|_{\beta} \leq |C|_{\beta}$. It follows by the valuation rules that for every ultimate β, $|B \rightarrow C|_{\beta} = 1$; so $||B \rightarrow C|| = 1$. Similarly, if $|B \rightarrow C|_{\Delta} = 0$ then $||B \rightarrow C|| = 0$. ■

> **Fundamental Theorem:** For any sentence A, $|A|_{\Delta} = ||A||$. That is, Δ is acceptable. (And since it was chosen to be bigger than an arbitrarily big τ, this gives that acceptable points occur arbitrarily late.)

Proof: Since Δ is ultimate, we know from Prop. 3 that if $||A|| = 1$, $|A|_{\Delta} = 1$, and similarly for 0. We need the converses; or equivalently, we need that if $||A|| = \frac{1}{2}$ then $|A|_{\Delta} = \frac{1}{2}$. Making the mini-stages explicit (and recalling that for any α, if a sentence has value $\frac{1}{2}$ at $\langle \alpha, \Omega \rangle$ then it has that value at all $\langle \alpha, \sigma \rangle$), the claim to be proved is that $(\forall A)(\forall \sigma)$(if $||A|| = \frac{1}{2}$ then $|A|_{\Delta,\sigma} = \frac{1}{2}$). Or reversing the quantifiers, that $(\forall \sigma)(\forall A)$(if $||A|| = \frac{1}{2}$ then $|A|_{\Delta,\sigma} = \frac{1}{2}$). Suppose this fails; let σ_0 be the smallest ordinal at which it fails. We get a contradiction by proving by induction on the complexity of A that

(*) $(\forall A)($ if $||A|| = \frac{1}{2}$ then $|A|_{\Delta,\sigma_0} = \frac{1}{2})$.

If A is a ground atomic sentence (atomic sentence of $\mathcal{L}$), $||A||$ is not $\frac{1}{2}$, so the claim is vacuous.

Similarly if A is $\mathsf{True}(t)$ where $den(t)$ is not a sentence.

Suppose A is $\mathsf{True}(t)$ where $den(t)$ is a sentence C. Then if $||A|| = \frac{1}{2}$, $||C|| = \frac{1}{2}$, since A and C must have the same value at each stage. So by choice of σ_0, $|C|_{\Delta,\sigma} = \frac{1}{2}$ for all $\sigma < \sigma_0$. But then by the valuation rules, $|\mathsf{True}(t)|_{\Delta,\sigma_0} = \frac{1}{2}$.

If A is a conditional, then by the valuation rules $|A|_{\Delta,\sigma_0}$ is $|A|_{\Delta,\Omega}$, i.e. $|A|_{\Delta}$, which is $\frac{1}{2}$ by Prop. 5.

The other cases use the claim that (*) holds for simpler sentences, and are fairly routine. E.g., if A is $\forall x A$, then if $||A|| = \frac{1}{2}$, there is a t_0 such that $||A(t_0/x)|| = \frac{1}{2}$ and for every t, $||A(t/x)|| \in \{\frac{1}{2}, 1\}$. But for any t for which $||A(t/x)||$ is $\frac{1}{2}$, including t_0, the induction hypothesis gives that $|A(t_0/x)|_{\Delta,\sigma_0} = \frac{1}{2}$; and for any t for which $||A(t/x)||$ is 1, $|A(t/x)|_{\Delta,\Omega}$ is 1 and so $|A(t/x)|_{\Delta,\sigma_0} \in \{\frac{1}{2}, 1\}$. So by the valuation rules for $\forall$, $|\forall x A|_{\Delta,\sigma_0} = \frac{1}{2}$. ■

Although I won't actually need this, it is helpful to remark that if Δ and $\Delta + \rho^{\star}$ are the first two acceptable ordinals, then for any β, $\Delta + \rho^{\star} \cdot \beta$ is acceptable. (The proof is by transfinite induction on β, and the successor case is trivial. For limits, use the continuity of conditionals at limits, and then extend to arbitrary sentences as in the proof of the Fundamental Theorem.) For sufficiently large β, $\Delta + \rho^{\star} \cdot \beta$ is itself of form $\rho^{\star} \cdot \beta$; letting $\Delta^{\star}$ be $\Delta + \rho^{\star} \cdot \beta$ for such a sufficiently large β, it follows that $\Delta^{\star} \cdot \gamma$ is acceptable for all non-zero γ. That is the form of the Fundamental Theorem promised early in Section 2.

4 The logic of $\rightarrow$

We need to define validity for sentences in $\mathcal{L}^+$. Probably the simplest approach is substitutional. Let Q be a pure quantificational language with the extra connective $\rightarrow$. (It can contain uninterpreted n-ary predicate letters for each n, including 0, and uninterpreted names.) Let a *realization* be a function assigning to each sentence letter of Q a sentence of $\mathcal{L}^+$ and to each n-ary predicate letter of Q a formula of $\mathcal{L}^+$ with exactly the first n variables of $\mathcal{L}^+$ free. Any realization s generates, for each formula C of Q, a formula C^s of $\mathcal{L}^+$ with the same free variables; if the n-ary predicate p is assigned the formula $\theta(v_1, ..., v_n)$, then any formula of form $p(v_{j_1}, ..., v_{j_n})$ will be assigned the corresponding $\theta(v_{j_1}, ..., v_{j_n})$. If C is a sentence of Q and Γ is a set of sentences of Q, call the inference from Γ to C *valid* iff for any such realization s such that $||\Gamma^s|| = 1$, $||C^s|| = 1$; here $||\Gamma^s||$ is the greatest lower bound of $\{A^s \mid A \in \Gamma\}$. (Note that s is a scheme for substituting sentences of $\mathcal{L}^+$, not $\mathcal{L}$; it is essential to consider substitutions of sentences containing 'True', so as to get sentences with value $\frac{1}{2}$.) And call C itself valid iff the inference from $\emptyset$ to C is.

Write $A_1, ..., A_n \models_{LCC} C$ to mean that the inference from $A_1, ..., A_n$ to C is valid. (LCC stands for "the logic of circularly defined concepts"–"circular logic" for short!–reflecting the view advocated in [6] that the problems about truth are simply instances of problems about circularly defined concepts generally.) I'll omit the subscript, except in a few contexts when classical validity $\models_{class}$ or Kleene validity $\models_K$ is also in view. Here is a partial axiomatization of the relation $\models_{LCC}$. (To avoid excess quantificational axioms, I imagine that $\exists$ is defined in terms of $\forall$ and $\neg$ in the usual way.)

Sentential Axioms:

A1 $\models A \rightarrow A$

A2 $\models \neg\neg A \rightarrow A$

A3a $\models A \rightarrow A \vee B$

A3b $\models B \rightarrow A \vee B$

A4a $\models A \wedge B \rightarrow A$

A4b $\models A \wedge B \rightarrow B$

A5 $\models A \wedge (B \vee C) \rightarrow (A \wedge B) \vee (A \wedge C)$

A6 $\models (A \rightarrow \neg B) \rightarrow (B \rightarrow \neg A)$

A7 $\models (A \rightarrow \neg A) \leftrightarrow \neg(\top \rightarrow A)$ [$\top$ is anything of form $B \rightarrow B$]

B1 $A, B \models A \wedge B$

B2 $A, A \rightarrow B \models B$ (Modus ponens)

B2* $A, \neg B \models \neg(A \rightarrow B)$

B3 $A \models B \rightarrow A$

B4 $A \rightarrow B \models (C \rightarrow A) \rightarrow (C \rightarrow B)$[6]

B4* $\neg[(C \rightarrow A) \rightarrow (C \rightarrow B)] \models \neg[A \rightarrow B]$

B5 $(A \rightarrow B) \wedge (A \rightarrow C) \models A \rightarrow (B \wedge C)$

B6 $(A \rightarrow C) \wedge (B \rightarrow C) \models (A \vee B) \rightarrow C$

Quantifier Axioms:

C1 $\models \forall x A \rightarrow A(x/t)$ [with the usual restrictions on legitimate substitution]

C2 $\models \forall x(A \vee Bx) \rightarrow A \vee \forall x Bx$, when x is not free in A

D1 $A(x) \models \forall x A(x)$

D2 $\forall x(Ax \rightarrow Bx) \models \forall x Ax \rightarrow \forall x Bx$

D3 $\forall x(\neg Ax \rightarrow Ax) \models \neg\forall x Ax \rightarrow \forall x Ax$

Structural Axiom:

$A \models A$

Rules: Aside from the two obvious structural rules

If $\Gamma \models A$ and $\Gamma \subseteq \Delta$ then $\Delta \models A$

If $\Gamma \models A$ and $\Gamma, A \models B$ then $\Gamma \models B$,

we need only disjunction elimination:

If $\Gamma, A \models C$ and $\Gamma, B \models C$ then $\Gamma, A \vee B \models C$.

Comparisons: There are at least two other systems in the literature that keep the full intersubstitutivity of $\mathsf{True}(\langle A \rangle)$ with A in a logic that contains a conditional validating $A \rightarrow A$ and thus in which we get the full truth schema: [2] and [5]. Those systems contain all the unstarred axioms of the present system except for B3 and A7;[7] they also contain "conditional strengthenings" of some of the B and D axioms: for instance, both contain the following strengthening of B4

$\models (A \rightarrow B) \wedge (C \rightarrow A) \rightarrow (C \rightarrow B)$,

[6]From this we can derive $A \rightarrow B \models (B \rightarrow C) \rightarrow (A \rightarrow C)$. To prove that, note first that the weaker form

(*) $A \rightarrow B, B \rightarrow C \models A \rightarrow C$

follows directly from B4 and B2, by relettering. Note second that $\models (A \rightarrow B) \rightarrow (\neg B \rightarrow \neg A)$ and $\models (\neg B \rightarrow \neg A) \rightarrow (A \rightarrow B)$. (Proof of first: A6 gives $(\neg B \rightarrow \neg B) \rightarrow (B \rightarrow \neg\neg B)$, hence $B \rightarrow \neg\neg B$ using A1 and B2. But then B4 gives $\models (A \rightarrow B) \rightarrow (A \rightarrow \neg\neg B)$. And A6 gives $\models (A \rightarrow \neg\neg B) \rightarrow (\neg B \rightarrow \neg A)$; so by (*), $\models (A \rightarrow B) \rightarrow (\neg B \rightarrow \neg A)$. Proof of second: Using A2 and B4 we get $\models (A \rightarrow \neg\neg B) \rightarrow (A \rightarrow B)$; and A6 gives $\models (\neg B \rightarrow \neg A) \rightarrow (A \rightarrow \neg\neg B)$; so by (*), $\models (\neg B \rightarrow \neg A) \rightarrow (A \rightarrow B)$.)

For the main result: $A \rightarrow B \models \neg B \rightarrow \neg A \models (\neg C \rightarrow \neg B) \rightarrow (\neg C \rightarrow \neg A)$, using the above (and B2) and B4. But also, $\models (\neg C \rightarrow \neg A) \rightarrow (A \rightarrow C)$ by the above, so $A \rightarrow B \models (\neg C \rightarrow \neg B) \rightarrow (A \rightarrow C)$ by (*). And $\models (B \rightarrow C) \rightarrow (\neg C \rightarrow \neg B)$ by the above, so $A \rightarrow B \models (B \rightarrow C) \rightarrow (A \rightarrow C)$ by (*).

[7]And [5] contains B4*.

and [5] contains the additional strengthening

$$\models (A \rightarrow B) \rightarrow [(C \rightarrow A) \rightarrow (C \rightarrow B)].$$

Despite the loss of those "conditional strengthenings" in the present system, I prefer this one because it contains B3: as we'll see, B3 is extremely important to shaping the approach to revenge problems that I'll be advocating. ([5] did contain a different weakening of the classical theorem $A \rightarrow (B \rightarrow A)$, but one rather less useful than B3; [2] contained nothing of this nature at all. Also, [5] did contain an important consequence of B3, the explosion rule; [2] doesn't have that either (it is a relevance logic).)

Soundness: It is straightforward to verify that these axioms are all true, and that the rules all preserve truth.[8] Let $\Gamma \vdash_{LCC} A$ mean that it is provable in this

[8]The validity of the main rule, disjunction elimination, is evident from the Fundamental Theorem: if $||A \vee B|| = 1$ then $|A \vee B|_\Delta = 1$, so at least one of $|A|_\Delta$ and $B|_\Delta$ is 1, so at least one of $||A||$ and $||B||$ is 1.

Here's a demonstration of the validity of most of the axioms. (The omitted ones are obvious.)

A1 through A5 and C1, C2: in each case the Kleene rules determine that the value of the antecedent is no greater than that of the consequent at any stage, so the conditional has value 1 at any stage other than 0.

A6: If $|A \rightarrow \neg B|_\alpha$ is 1 then for all sufficiently big predecessors β of α, $|A|_\beta \leq 1 - |B|_\beta$ and hence $|B|_\beta \leq 1 - |A|_\beta$, so $|B \rightarrow \neg A|_\alpha$ is 1. Similarly, if $|B \rightarrow \neg A|_\alpha$ is 0, $|A \rightarrow \neg B|_\alpha$ is 0. So for all α, $|A \rightarrow \neg B|_\alpha \leq |B \rightarrow \neg A|_\alpha$, and so $|(A \rightarrow \neg B) \rightarrow (B \rightarrow \neg A)|_\alpha$ is 1 whenever $\alpha > 0$.

A7: The left and right hand sides have the same value for each α.

B2: Suppose $||A||$ and $||A \rightarrow B||$ are both 1. Then there are α_1 and α_2 such that $(\forall \beta \geq \alpha_1)(|A|_\beta = 1)$ and $(\forall \beta \geq \alpha_2)(|A \rightarrow B|_\beta = 1)$. The latter implies that $(\forall \beta \geq \alpha_2 + 1)(|A|_\beta \leq |B|_\beta)$. So letting α be $max\{\alpha_1, \alpha_2 + 1\}$, it follows that $(\forall \beta \geq \alpha)(|B|_\beta = 1)$; so $||B|| = 1$.

$B2^\star$: If $||A||$ and $||\neg B||$ are both 1, there's an α such that $(\forall \beta \geq \alpha)[|A|_\beta = 1 \wedge |B|_\beta = 0]$; so $(\forall \beta \geq \alpha + 1)[|A \rightarrow B|_\beta = 0]$, so $(\forall \beta \geq \alpha + 1)[|\neg(A \rightarrow B)|_\beta = 1]$, so $||\neg(A \rightarrow B)|| = 1$.

B3: Suppose $||A|| = 1$. Then for some α, $(\forall \beta \geq \alpha)(|A|_\beta = 1)$. So $(\forall \beta \geq \alpha + 1)(|B \rightarrow A|_\beta = 1)$, so $||B \rightarrow A|| = 1$.

B4: Suppose $||A \rightarrow B|| = 1$. Then there is an α such that for any $\beta \geq \alpha$, $|A|_\beta \leq |B|_\beta$. Now suppose $||(C \rightarrow A) \rightarrow (C \rightarrow B)|| < 1$. Then for some $\beta_0 \geq \alpha + 1$, $|C \rightarrow A|_{\beta_0} > |C \rightarrow B|_{\beta_0}$; choose the smallest. β_0 can't be a successor: that would require $|C|_{\beta_0 - 1} \leq |A|_{\beta_0 - 1}$ and $|C|_{\beta_0 - 1} > |B|_{\beta_0 - 1}$, contrary to the fact that $|A|_{\beta_0 - 1} \leq |B|_{\beta_0 - 1}$. For the case where β_0 is a limit, we have that either $|C \rightarrow A|_{\beta_0} = 1$ or $|C \rightarrow B|_{\beta_0} = 0$. In the first case, $C \rightarrow A$ eventually has value 1 prior to β_0, so by definition of β_0, $C \rightarrow B$ does too, so $|C \rightarrow B|_{\beta_0} = 1$. In the second case, $C \rightarrow B$ eventually has value $\neq 1$ prior to β_0, so by definition of β_0, $C \rightarrow A$ does too, so $|C \rightarrow A|_{\beta_0} = 0$. Contradiction.

$B4^\star$: Suppose $||\neg((C \rightarrow A) \rightarrow (C \rightarrow B))|| = 1$, i.e. $||(C \rightarrow A) \rightarrow (C \rightarrow B)|| = 0$. Then there is an α such that whenever $\beta \geq \alpha$, $|C \rightarrow A|_\beta > |C \rightarrow B|_\beta$. In particular this is so for all β of form $\gamma + 1$ (where values of $\frac{1}{2}$ don't occur for conditionals), so we have that for all sufficiently large γ, $|C|_\gamma \leq |A|_\gamma$ and $|C|_\gamma > |B|_\gamma$, hence $|A|_\gamma > |B|_\gamma$. So for all sufficiently large γ, $|A \rightarrow B|_{\gamma+1} = 0$, and by the continuity of conditionals at limits this implies that $|\neg(A \rightarrow B)|_\delta = 1$ for all sufficiently large δ, and hence $||\neg(A \rightarrow B)|| = 1$.

B5: Suppose $||(A \rightarrow B) \wedge (A \rightarrow C)||$ is 1, i.e. $||A \rightarrow B|| = ||A \rightarrow C|| = 1$. Then for all sufficiently large α, $|A|_\alpha \leq |B|_\alpha$ and for all sufficiently large α, $|A|_\alpha \leq |C|_\alpha$; so for all sufficiently large α, $|A|_\alpha \leq |B \wedge C|_\alpha$; so $||A \rightarrow (B \wedge C)|| = 1$. (B6 is similar.)

D2: Suppose $||\forall x(Ax \rightarrow Bx)|| = 1$. Then there is an α such that $(\forall \beta \geq \alpha)(|\forall x(Ax \rightarrow Bx)|_\beta = 1)$, hence $(\forall \beta \geq \alpha)(\forall t)(|At \rightarrow Bt|_\beta = 1)$, hence $(\forall \beta \geq \alpha)(\forall t)(|At|_\beta \leq |Bt|_\beta)$. Suppose $||\forall xAx \rightarrow \forall xBx|| < 1$. Then for some $\beta \geq \alpha + 1$, $|\forall xAx \rightarrow \forall xBx|_\beta < 1$, so for some $\beta \geq \alpha$, $|\forall xAx|_\beta > |\forall xBx|_\beta$; this is clearly incompatible with the above given that everything has a name.

D3: Suppose $||\forall x(\neg Ax \rightarrow Ax)|| = 1$. Then for all sufficiently large α, $|\forall x(\neg Ax \rightarrow Ax)|_\alpha = 1$, so for all sufficiently large α and all t, $|\neg At \rightarrow At|_\alpha = 1$, so for all t and all sufficiently large α, $|At|_\alpha \geq \frac{1}{2}$. So for all sufficiently large α, $|\forall xAx|_\alpha \geq \frac{1}{2}$, and so for all sufficiently large α, $|\neg \forall xAx|_\alpha \leq |\forall xAx|_\alpha$, and so $||\neg \forall xAx \rightarrow \forall xAx|| = 1$.

system that $\Gamma \models A$. So we have that if $\Gamma \vdash_{LCC} A$ then $\Gamma \models A$.

We can easily derive the deMorgan laws, in the strong form $\vdash_{LCC} \neg(A \vee B) \rightarrow \neg A \wedge \neg B$ and its converse, and their analogues with $\wedge$ and $\vee$ switched; also, the converse of A2.

And we have the following obvious metatheorem:

> **General Substitutivity Rule:** Let A and B be any formulas with the same free variables, let Ψ_A be any formula that contains A as a subformula, and let Ψ_B result from Ψ_A by substituting B for any number of occurrences of A. Then $A \leftrightarrow B \vdash_{LCC} \Psi_A \leftrightarrow \Psi_B$. Indeed, if all substituted occurrences of A in Ψ_A are positive then $A \rightarrow B \vdash_{LCC} \Psi_A \rightarrow \Psi_B$; and if all are negative then $A \rightarrow B \vdash_{LCC} \Psi_B \rightarrow \Psi_A$.[9]

(Given this, the intersubstitutivity of $\mathsf{True}(\langle A \rangle)$ with A follows from the truth schema; this seems of interest, even though the construction yields the intersubstitutivity of $\mathsf{True}(\langle A \rangle)$ with A more directly.)

The significance of $\rightarrow$ in this system is partially clarified by the following theorem, three clauses of which require B3:

Theorem on $\rightarrow$ and $\supset$: Let $A \supset B$ abbreviate $\neg A \vee B$. Then

(ia) $A \supset B \vdash_{LCC} A \rightarrow B$;

(ib) $(A \vee \neg A) \wedge (B \vee \neg B) \vdash_{LCC} (A \supset B) \rightarrow (A \rightarrow B)$;

(iia) $(A \vee \neg A) \wedge (A \rightarrow B) \vdash_{LCC} A \supset B$;

(iib) $(A \vee \neg A) \wedge (B \vee \neg B) \vdash_{LCC} (A \rightarrow B) \rightarrow (A \supset B)$

(If it seems odd that the excluded middle premises are required for (ib) but not for (ia), a glance at the value-table given at the end of Section 2 may help.)

Proof: It's useful first to note that B3 together with other rules yields

B3$^{\#}$: $\neg A \vdash_{LCC} A \rightarrow B$.

For B3 gives $\neg A \vdash_{LCC} \neg B \rightarrow \neg A$; which with A6 and B2 gives $\neg A \vdash_{LCC} A \rightarrow \neg\neg B$; but by B1 and B4 we get $\vdash (A \rightarrow \neg\neg B) \rightarrow (A \rightarrow B)$, so by B2 we get $\neg A \vdash_{LCC} A \rightarrow B$.

[9] Proof: It suffices to consider a single substituted occurrence, and we prove the result for this case by induction on the complexity of the embedding of that occurrence. Trivial for the case where Ψ_A is A. Suppose we've established that if A is positive in θ_A then $A \rightarrow B \vdash_{LCC} \theta_A \rightarrow \theta_B$, and if negative then $A \rightarrow B \vdash_{LCC} \theta_B \rightarrow \theta_A$; it follows that if A is positive in $\neg\theta_A$ and hence negative in θ_A then $A \rightarrow B \vdash_{LCC} \theta_B \rightarrow \theta_A$ and hence $A \rightarrow B \vdash_{LCC} \neg\theta_A \rightarrow \neg\theta_B$; similarly when A is negative in θ_A. If A is positive in $\theta_A \wedge C$ by being positive in the first conjunct: $A \rightarrow B \vdash_{LCC} \theta_A \rightarrow \theta_B$, so $A \rightarrow B \vdash_{LCC} \theta_A \wedge C \rightarrow \theta_B$ using A4a and B4 (and B2); and $\vdash_{LCC} \theta_A \wedge C \rightarrow C$, so $A \rightarrow B \vdash_{LCC} \theta_A \wedge C \rightarrow \theta_B \wedge C$ using B5. If A is negative in $\theta_A \wedge C$ by being negative in the first conjunct: $A \rightarrow B \vdash_{LCC} \theta_B \rightarrow \theta_A$, so by an analogous argument $A \rightarrow B \vdash_{LCC} \theta_B \wedge C \rightarrow \theta_A \wedge C$. Disjunction (and second conjunct or disjunct) is similar. If A is positive in $\forall x \theta_A(x)$ and hence in $\theta_A(x)$: $A \rightarrow B \vdash_{LCC} \theta_A(x) \rightarrow \theta_B(x)$, so $A \rightarrow B \vdash_{LCC} \forall x \theta_A(x) \rightarrow \forall x \theta_B(x)$ using D1 and C2. (Negative similar.) If A is positive in $C \rightarrow \theta_A$ by being positive in the consequent, then $A \rightarrow B \vdash_{LCC} \theta_A \rightarrow \theta_B$, so $A \rightarrow B \vdash_{LCC} (C \rightarrow \theta_A) \rightarrow (C \rightarrow \theta_B)$ by B4; if A is positive in $\theta_A \rightarrow C$ by being negative in θ_A, then $A \rightarrow B \vdash_{LCC} \theta_B \rightarrow \theta_A$, so $A \rightarrow B \vdash_{LCC} (\theta_A \rightarrow C) \rightarrow (\theta_B \rightarrow C)$ by the analog of B4 proved in note 4. (Negative similar.)

Proof of (ia): B3 and its consequence B3$^{\#}$ give both that $B \vdash_{LCC} A \rightarrow B$ and that $\neg A \vdash_{LCC} A \rightarrow B$; so $\neg A \vee B \vdash_{LCC} A \rightarrow B$ by $\vee$-Elimination.

Proof of (ib): Applying B3 to the result of (ia), we get $\neg A \vee B \vdash_{LCC} (A \supset B) \rightarrow (A \rightarrow B)$. Also $A \wedge \neg B \vdash_{LCC} \neg(A \supset B)$, so by B3$^{\#}$, $A \wedge \neg B \vdash_{LCC} (A \supset B) \rightarrow (A \rightarrow B)$. So by $\vee$-elimination, $\neg A \vee B \vee (A \wedge \neg B) \vdash_{LCC} (A \supset B) \rightarrow (A \rightarrow B)$; and since $(A \vee \neg A) \wedge (B \vee \neg B) \vdash_{LCC} \neg A \vee B \vee (A \wedge \neg B)$, the result follows.

Proof of (iia): By distributivity, the premise is equivalent to $[A \wedge (A \rightarrow B)] \vee [\neg A \wedge (A \rightarrow B)]$. But $A \wedge (A \rightarrow B) \vdash_{LCC} B \vdash_{LCC} A \supset B$, and $\neg A \wedge (A \rightarrow B) \vdash_{LCC} \neg A \vdash_{LCC} A \supset B$. So (iia) holds, by $\vee$-Elim.

Proof of (iib): By B3, $A \supset B \vdash (A \rightarrow B) \rightarrow (A \supset B)$. And $A \wedge \neg B \vdash \neg(A \rightarrow B)$ by B2*, so $A \wedge \neg B \vdash (A \rightarrow B) \supset (A \supset B)$ by B3$^{\#}$. So by $\vee$-Elim, $\neg A \vee B \vee (A \wedge \neg B) \vdash (A \supset B) \rightarrow (A \rightarrow B)$; the result follows as in (ib). ■

The upshot of (ib) and (iib) is that there's no difference between $\rightarrow$ and $\supset$ in contexts, like ordinary arithmetic, in which excluded middle is assumed to hold. (This fact marks a big difference between the conditional used here and that in [5]; only (iia) of the theorem is valid for that conditional.)[10]. More precisely:

Corollary on $\rightarrow$ and $\supset$: For any formula A of $\mathcal{L}^{+}$, let LEM_A be the universal closure of $A \vee \neg A$. And for any formula C of $\mathcal{L}^{+}$, let $AtLEM_C$ be $\{LEM_A \mid A$ is an atomic subformula of $C\}$. (For present purposes, $A(t)$ does not count as a subformula of $\forall x A(x)$ for terms t other than x.) Also, let $C^{\star}$ be the result of replacing all occurrences of $\rightarrow$ by $\supset$ (and, if you like, translating the result into the official language of $\neg$, $\vee$ etc.). Then: $AtLEM_C \vdash_{LCC} C \leftrightarrow C^{\star}$ (where this abbreviates $(C \rightarrow C^{\star}) \wedge (C^{\star} \rightarrow C)$).

Proof: From (ib) and (iib) of the Theorem and the General Substitutivity Rule, we clearly get that $\{AtLEM_A \mid A$ is a subformula of $C\} \vdash_{LCC} C \leftrightarrow C^{\star}$. It remains only to show that LEM for the *atomic* subformulas suffices to get LEM for *all* subformulas, that is

Lemma: For any C, $AtLEM_C$ entails LEM_A for any A that is a subformula of C.

Proof in footnote.[11] ■

[10]The quasi-semantics of [5] involved an interpretation of $\rightarrow$ in terms of derivability. When A is itself an arithmetical theorem, say $0 = 0$, $A \rightarrow B$ becomes in effect $Provable(\langle B \rangle)$; taking B to be the Gödel sentence invalidates (ia), (ib) and (iib). [(iib) becomes in effect $Provable(\langle Provable(\langle G \rangle) \supset G \rangle)$, which is equivalent to $Provable(\langle G \rangle)$ via either Löb's theorem or the equivalence of G to $\neg Provable(\langle G \rangle)$.] (iia) is still valid on that semantics: given excluded middle, $A \rightarrow B$ reduces on the semantics to $\Box(A \supset B)$ where $\Box$ is the provability operator of GLS rather than of G. [For GLS and G, see [1]]

[11]Proof sketch: The claim is trivial if A is atomic. For negation, we must merely get from $\forall x_1 ... \forall x_k (A \vee \neg A)$ to $\forall x_1 ... \forall x_k (\neg A \vee \neg\neg A)$; but passing from $A \vee \neg A$ to $\neg A \vee \neg\neg A$ is elementary, and the quantifier rules allow the addition of the string $\forall x_1 ... \forall x_k$ to premises and conclusion. The conjunction and disjunction cases are similarly easy. For universal quantification, we must get from $\forall x_1 ... \forall x_k (A \vee \neg A)$ to $\forall x_1 ... \forall x_{k-1} (\forall x_k A \vee \neg \forall x_k A)$; for this it suffices to get from $\forall x_k (A \vee \neg A)$ to $\forall x_k A \vee \neg \forall x_k A$. But we can get from $A \vee \neg A$ to $A \vee \neg \forall x_k A$ using C1 and $\vee$-Elim; so we can get from $\forall x_k (A \vee \neg A)$ to $\forall x_k A \vee \neg \forall x_k A$ by the additional use of C2. For $\rightarrow$, we must get from $A \vee \neg A$ and $B \vee \neg B$ to $(A \rightarrow B) \vee \neg (A \rightarrow B)$; we can then add quantifiers to premise and conclusion as above. But $A \wedge \neg B \vdash \neg(A \rightarrow B)$ by B2*, and $A \supset B \vdash A \rightarrow B$ by theorem above, and $(A \vee \neg A) \wedge (B \vee \neg B) \vdash (A \wedge \neg B) \vee (A \supset B)$; so $(A \vee \neg A) \wedge (B \vee \neg B) \vdash (A \rightarrow B) \vee \neg (A \rightarrow B)$.

Another important consequence of the theorem on $\rightarrow$ and $\supset$ (this time requiring only (ia) of the theorem) is:

Corollary on Explosion: $A \wedge \neg A \vdash_{LCC} B$.

Proof: $A \vee B \vdash_{LCC} \neg A \rightarrow B$, by (ia) of the theorem. From this and B2, we get $\neg A, A \vee B \vdash_{LCC} B$; using A3, we get $\neg A, A \vdash_{LCC} B$ (which yields the above via B1). ■

Given this, it is easy to verify that the system can derive any Kleene-valid inference (if $\Gamma \models_K A$ then $\Gamma \vdash_{LCC} B$); where an inference is Kleene-valid if whichever of the values 0, $\frac{1}{2}$, 1 is assigned to each atomic sentence *or conditional sentence*, the usual strong Kleene rules (taking quantifiers as substitutional) give B the value 1 whenever they give each member of Γ the value 1.

And from this and the above we get that full classical reasoning, *including the treatment of* $\rightarrow$ *as* $\supset$, is available in any context in which excluded middle can be assumed. More precisely, if Γ is any set of formulas of $\mathcal{L}^+$, let $AtLEM_\Gamma$ be $\{LEM_A \mid A$ is atomic and occurs in some member of $\Gamma\}$ and let $\Gamma^\star$ be $\{C^\star \mid C \in \Gamma\}$, where LEM_A and $C^\star$ are as defined in the Corollary on $\rightarrow$ and $\supset$. Then

Corollary on relation to classical logic: For any formula B of $\mathcal{L}^+$ and any set Γ of such formulas:

If $\Gamma^\star \models_{class} B^\star$, then $\Gamma \cup AtLEM_{\Gamma \cup \{B\}} \vdash_{LCC} B$.

Proof: $\Gamma^\star$ is in the $\rightarrow$-free language, and $\Gamma^\star \models_{class} B^\star$, so $\Gamma^\star \cup AtLEM_\Gamma \vdash_K B^\star$ (since as is well-known, Kleene logic plus excluded middle for the relevant vocabulary yields all classically valid inferences in that vocabulary, and since the atomic subformulas of $\Gamma^\star$ are the same as those of Γ). So by the observation after the corollary on Explosion, $\Gamma^\star \cup AtLEM_\Gamma \vdash_{LCC} B^\star$. So $\Gamma^\star \cup AtLEM_{\Gamma \cup \{B\}} \vdash_{LCC} B^\star \cup AtLEM_B$ (i.e. each member of the consequent set is derivable from the antecedent set). But by the corollary on $\rightarrow$ and $\supset$, we have both that $\Gamma \cup AtLEM_{\Gamma \cup \{B\}} \vdash_{LCC} \Gamma^\star \cup AtLEM_{\Gamma \cup \{B\}}$ and that $B^\star \cup AtLEM_B \vdash_{LCC} B$; putting all these together, we get that $\Gamma \cup AtLEM_{\Gamma \cup \{B\}} \vdash_{LCC} B$. ■

The import of all of this is that we can take LCC as our general background logic, and simply add instances of excluded middle as "non-logical premises" wherever it seems appropriate–for instance, in arithmetic, in physics, in set theory. This will legitimize full classical reasoning in those areas, *including the treatment of* $\rightarrow$ *as* $\supset$. Thus *there is no need to worry that using LCC as the general background logic will cripple reasoning in any domain where excluded middle is legitimate.*[12]

[12] A small caution about the application of this to theories that contain axiom schemas, for instance, Peano arithmetic or ZFC: the assumption of excluded middle for the basic arithmetic or set-theoretic vocabulary does not suffice to ensure classical reasoning with regard to instances of the schemas that involve vocabulary outside of the arithmetic or set-theoretic vocabulary. (Consider the set whose only member is 1 if the Liar is true and whose only member is 0 otherwise!)

A related point: some care is needed about the extension of arithmetic or set theory to include a truth predicate. For instance, in the arithmetic case, if we want to allow mathematical induction on formulas containing 'True' (and we *should* want to!), then induction needs to be put as a rule

Some non-laws: Certain laws that are valid for the classical '$\supset$' (and hence valid for '$\rightarrow$' in the context of excluded middle) fail for '$\rightarrow$' when excluded middle is not presupposed. In many cases, their failure is virtually inevitable: they will fail in any remotely reasonable system that yields the truth schema. Three laws that are inevitably invalid in this sense are:

Importation: $A \rightarrow (B \rightarrow C) \models_? A \wedge B \rightarrow C$

Contraction: $A \rightarrow (A \rightarrow B) \models_? A \rightarrow B$

$\rightarrow$-Introduction: From $A \models B$ infer$_?$ $\models A \rightarrow B$.

That Contraction fails is virtually inevitable if we are to save the truth schema, due to the Curry Sentence: a sentence K which is equivalent to $\mathsf{True}(\langle K \rangle) \rightarrow \perp$, where $\perp$ is an absurdity such as $0 = 1$. For one can derive $\perp$ from $\mathsf{True}(\langle K \rangle) \leftrightarrow K$ using only very uncontroversial axioms plus Contraction; that's the Curry Paradox. (For discussion, see [5].) As noted in the next section, the values of K are as follows:

$$|K|_\alpha = \begin{cases} \frac{1}{2} & \text{whenever } \alpha \text{ is 0 or a limit} \\ 0 & \text{whenever } \alpha \text{ is odd} \\ 1 & \text{whenever } \alpha \text{ is an even successor} \end{cases}$$

Given this, it is easily checked that for each α, $|K \rightarrow \perp|_\alpha = |K|_\alpha$, from which it follows that $||K \rightarrow (K \rightarrow \perp)||$ is 1 and $||K \rightarrow \perp||$ is $\frac{1}{2}$.

Taking A and B both to be K and C to be $\perp$ also gives a counterexample to Importation; not surprising, since Contraction is virtually identical to the special case of Importation with B set equal to A. (The initial impression of the complete obviousness of Contraction rests largely on the assumption of Importation.) And taking A to be K and B to be $\perp$ gives a counterexample to $\rightarrow$-introduction; again not surprising, both because of a famous alternate form of the Curry Paradox using $\rightarrow$-introduction instead of Contraction and because Contraction is easily derivable using two applications of modus ponens followed by an $\rightarrow$-introduction. (Insofar as the initial impression of the complete obviousness of Contraction *doesn't* rest on the assumption of Importation, it probably rests on this derivation.)

K in fact can be used to provide counterexamples to a number of other classically-valid rules. For instance,

Permutation: $A \rightarrow (B \rightarrow C) \models_? B \rightarrow (A \rightarrow C)$

Weak-Perm: $\models_? (\top \rightarrow C) \rightarrow C$ [where $\top$ is $A \rightarrow A$, or $0 = 0$]

---

rather than an axiom. We can use the strong rule

$F(0) \wedge \forall n[F(n) \rightarrow F(n+1)] \models \forall n F(n)$

rather than the weaker rule with '$\supset$' in place of '$\rightarrow$'. But we can then derive the induction *axiom* for predicates for which excluded middle holds, and we can replace $\rightarrow$ with $\supset$ there; so we get all of standard number theory. (Indeed, we can prove more besides, since an induction on the truth of theorems is now available in the theory, which will enable us to prove the Gödel sentence of the 'True'-free theory. But despite the availability of inductions involving 'True', we will not be able to prove the Gödel sentence of the full theory containing 'True': the lack of excluded middle will prevent this.)

I call the latter Weak Permutation since a special case of Permutation is $(\top \rightarrow C) \rightarrow (\top \rightarrow C) \models_? \top \rightarrow ((\top \rightarrow C) \rightarrow C)$; by A1 this yields $\models_? \top \rightarrow ((\top \rightarrow C) \rightarrow C)$, and by B2 that yields Weak Permutation. So it suffices to get a counterexample to Weak Permutation, and for this we just let C be the Curry sentence K. ($\top \rightarrow K$ is equivalent to $\neg K$, so when α is odd $|\top \rightarrow K|_\alpha$ is 1 and $|K|_\alpha$ is 0, so when α is an even successor $|(\top \rightarrow K) \rightarrow K|_\alpha$ is 0, so $||(\top \rightarrow K) \rightarrow K|| \neq 1$.) Again, Permutation may seem completely obvious, but I think that that impression rests on the assumption of Importation, and the additional assumption of its converse, Exportation.

Exportation ($A \wedge B \rightarrow C \models_? A \rightarrow (B \rightarrow C)$) is invalid for similar reasons. Here, let A and C be the Curry sentence K and B to be $\top$; the premise $K \wedge \top \rightarrow K$ is obviously valid, but the consequent $K \rightarrow (\top \rightarrow K)$ is not: it's value at stage α is 0 whenever α is odd and neither 1 nor the successor of a limit, so $||K \rightarrow (\top \rightarrow K)|| \neq 1$.

In addition, consider the "conditional strengthenings" of all the B and D rules other than B1 and D1. (I exempt D1 because in that case the failure of the strengthening is unsurprising.) The strengthening of B3 would be $\models_? A \rightarrow (B \rightarrow A)$. Take A to be the Curry sentence K and B to be $\top$. The instance in question, then, is $\models_? K \rightarrow (\top \rightarrow K)$, which we've just seen is invalid.

Indeed, the "conditional strengthening" of B3 fails even in the special case $(A \rightarrow C) \rightarrow (A \rightarrow (A \rightarrow C))$: Take A to be $\top$, C to be K. Since $\top \rightarrow K$ is effectively equivalent to K, this reduces to the previous counterexample.

The strengthening of B4 (and B4*) would be $\models_? (A \rightarrow B) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C))$. Take A to be K and B and C both $\bot$. We get $\models_? (K \rightarrow \bot) \rightarrow ((\bot \rightarrow \bot) \rightarrow (K \rightarrow \bot))$, which is in effect $\models_? (K \rightarrow \bot) \rightarrow (\top \rightarrow (K \rightarrow \bot))$. Since as remarked above $|K \rightarrow \bot|_\alpha$ is always the same as $|K|_\alpha$, this is effectively equivalent to $\models_? K \rightarrow (\top \rightarrow K)$, which we've already seen to fail.

An alternative strengthening of B4 would be $\models_? (A \rightarrow B) \wedge (B \rightarrow C) \rightarrow (A \rightarrow C)$. Let A be $\top$, B be K, C be $\bot$. We get $\models_? (\top \rightarrow K) \wedge (K \rightarrow \bot) \rightarrow (\top \rightarrow \bot)$. At any limit ordinal, the antecedent gets value $\frac{1}{2}$ and the consequent value 0, so the conditional has value 0 at every successor of a limit and consequently can't have ultimate value 1.

The strengthening of the B5 rule would be $\models_? ((A \rightarrow B) \wedge (A \rightarrow C)) \rightarrow (A \rightarrow (B \wedge C))$. Let A be $\top$, B be K and C be $\top \rightarrow K$. Then $B \wedge C$ is $\frac{1}{2}$ at 0 and limits, 0 otherwise, so $A \rightarrow (B \wedge C)$ is 0 at stages above 0; but $A \rightarrow B$ and $A \rightarrow C$ are both $\frac{1}{2}$ at all limits. The strengthening of B6 is similarly invalid.

The most obvious strengthening of B2 is $\models_? (A \wedge (A \rightarrow B)) \rightarrow B$; taking A to be $\top$, this is in effect weak permutation, and fails when B is K.

The failure of the strengthening of B4 entails failure of the corresponding strengthening of D2 ($\models_? \forall x(Ax \rightarrow Bx) \rightarrow (\forall xAx \rightarrow \forall xBx)$). The strengthening of D3 ($\models_? \forall x(\neg Ax \rightarrow Ax) \rightarrow (\neg \forall xAx \rightarrow \forall xAx)$) also fails. For let $A(t)$ always be either K or $\neg K$, with some of each. Then $\forall xAx$ has value 0 at successors, $\frac{1}{2}$ at 0 and limits, and so $\neg \forall xAx \rightarrow \forall xAx$ has value 0 at limits not divisible by ω^2. But for each t, $\neg At \rightarrow At$ alternates between 0 and 1 at successors, so is $\frac{1}{2}$ at all limits, so the same is true of its universal quantification. Since there

are arbitrarily high limits not divisible by ω^2, there are arbitrarily high limits at which the value of the antecedent is greater than that of the consequent.

Another rule whose failure is almost inevitable–though in this case the argument for the failure requires the Explosion rule $A \wedge \neg A \models B$, which as we've seen is a consequence of B3—is

Pseudo-reductio: $A \rightarrow \neg A \models_? \neg A$.

(Axiom A7 does yield a weaker version of this.) In any reasonable system with Explosion, Pseudo-Reductio must fail because of the Liar paradox (if the system is not to be trivial in the sense of having every sentence as a theorem). For if L_0 is a sentence that asserts its own untruth, $L_0 \rightarrow \neg L_0$ and $\neg L_0 \rightarrow L_0$ must be theorems in any theory that includes the Tarski biconditionals; but the latter together with $\neg L_0$ leads to $L_0 \wedge \neg L_0$ by B2 and B1, and hence to any absurdity you like; so the inference from $L_0 \rightarrow \neg L_0$ to $\neg L_0$ must fail, if triviality is to be avoided. As noted in the next section, the Liar sentence has value $\frac{1}{2}$ at each level, which gives $L_0 \rightarrow \neg L_0$ the value 1 at levels > 0 while $\neg L_0$ has value $\frac{1}{2}$.

The Liar sentence also shows the invalidity of

$A \rightarrow \neg B \models_? \neg(A \rightarrow B)$

and

$\neg(A \rightarrow B) \models_? A \rightarrow \neg B$:

in the first case, take A and B to be L_0, and in the second case take A to be $\top$ and B to be L_0.

5 Transfinite hierarchies of paradoxes: a first step toward defusing the "revenge problem"

The most famous paradoxical sentence (and one that is particularly easy to deal with since it doesn't involve the $\rightarrow$) is the Liar sentence L_0, which in effect asserts its own untruth. Officially, it has the form $\exists w[\mathsf{SA}(w, \mathbf{k}) \wedge \neg \mathsf{True}(w)]$, where $\mathbf{k}$ is the standard name of $\exists w[\mathsf{SA}(w, v_1) \wedge \neg \mathsf{True}(w)]$ and SA is the arithmetical formula for "self application". But for all α, $|\mathsf{SA}(t, \mathbf{k})|_\alpha$ is 1 if $den(t)$ is the Gödel code of L_0 and 0 otherwise, so by the Kleene rules, $|L_0|_\alpha$ is $1 - |\mathsf{True}(\langle L_0 \rangle)|_\alpha$ for each α. Since $|\mathsf{True}(\langle A \rangle)|_\alpha$ always has the same value as $|A|_\alpha$, for any A and α, we have that for all α, $|L_0|_\alpha$ is $1 - |L_0|_\alpha$; i.e., for all α, $|L_0|_\alpha$ is $\frac{1}{2}$ (and hence $|\mathsf{True}(\langle L_0 \rangle)|_\alpha$ and $|\neg\mathsf{True}(\langle L_0 \rangle)|_\alpha$ are $\frac{1}{2}$ too).

Let's also consider the Truth-teller sentence T_0, which attributes truth to itself. The Tarski truth schema would be satisfied by assigning any of the three semantic values to it. But the semantics dictates that it gets value $\frac{1}{2}$: for neither it nor its negation are in the minimal fixed point at level 0, and since the sentence doesn't contain an $\rightarrow$ its value is not affected by later stages.

Presumably it is incorrect to assert sentences with values less than 1, so it is incorrect to assert that the Liar is true, *and also incorrect to assert that it is not*

true. One should not object that it must be either true or not true, and therefore must be correct to assert one or the other; for it is incorrect to assert that it is either true or not true. (Similarly for the Truth-teller.) Let's define 'false' as 'has a true negation', and 'lacks truth value' as 'is neither true nor false', and 'has at most one truth value' as 'is not both true and false'. Then it is also incorrect to assert of either the Liar or the Truth-teller that it lacks truth value; or that it has truth value; or that it either has or lacks truth value. (Note this well: this is *not* a theory that postulates "truth value gaps".)[13] It is likewise incorrect to assert of either the Liar or the Truth teller that it has at most one truth value, or that it doesn't have at most one, or that it either has at most one or doesn't have at most one.

Does this mean that the "singular" status of the Liar and the Truth-Teller can't be stated in the language (but only in the set-theoretic metalanguage where we talk about semantic values)? No: we can capture it perfectly well within the language, by defining a "determinately operator". The definition I propose is: DA abbreviates $(\top \rightarrow A) \wedge A$.

Note that B3 gives the inference rule $A \vdash DA$, and that the converse holds in the strengthened form $\vdash DA \rightarrow A$. (If all we wanted was the unstrengthed form, there'd have been no reason to build the 2^{nd} conjunct into the definition of DA.)

This operator will be of much concern in what follows, so let's note the values it takes on at (super-)stages. Applying the rules for $\rightarrow$, we get:

$$|DA|_\alpha = \begin{cases} 1 & |A|_\alpha = 1 \textit{ and } (\exists\beta < \alpha)(\forall\gamma \in [\beta, \alpha))(|A|_\gamma = 1) \\ 0 & |A|_\alpha = 0 \textit{ or } (\exists\beta < \alpha)(\forall\gamma \in [\beta, \alpha))(|A|_\gamma \neq 1) \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

As special cases we get:

$$|DA|_0 = \begin{cases} 0 & \text{if } |A|_0 = 0 \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

$$|DA|_{\beta+1} = \begin{cases} |A|_{\beta+1} & \text{if } |A|_\beta = 1 \\ 0 & \text{otherwise; that is, 0 iff either } |A|_{\beta+1} = 0 \text{ or } |A|_\beta \neq 1 \end{cases}$$

Observations:

(a) for all α, $|DA|_\alpha \leq |A|_\alpha$;

(b) *If A is a conditional*, then $|DA|_\lambda = 1$ iff $|A|_\lambda = 1$ (where λ is a limit);

(c) For *all* A, $|DA|_\Delta = 1$ iff $|A|_\Delta = 1$ (where Δ is an acceptable point);

(d) If there is an $\alpha < \beta$ such that for all γ in $[\alpha, \beta]$, $|A|_\gamma \leq |B|_\gamma$, then $|DA|_\beta \leq |DB|_\beta$;

(e) For all α, $|D(A \wedge B)|_\alpha \leq min\{|DA|_\alpha, |DB|_\alpha\}$

[13] I have already cautioned against equating having semantic value $\frac{1}{2}$ with being neither true nor false. If you think that the fact that sentences like the Liar get value $\frac{1}{2}$ in the formal semantics means that we are nonetheless *in some sense* postulating a truth value gap, read on: I believe that the rest of the paper undermines this (so far very vague) suggestion.

The left to right of (b) and (c) come from (a); the valuation rules for $\rightarrow$ yield the right to left of (b). And the right hand side of (c) implies that $||A|| = 1$ by the Fundamental Theorem, which implies that $||DA|| = 1$, which implies that $|DA|_\Delta$ is 1. (d) is evident from the valuation rules for D. (e) follows from (d) when $\alpha > 0$, and is evident (even in strengthened equality form) for $\alpha = 0$. ■

Note that even at Δ, we needn't have equality in (e): $|D(A \wedge B)|_\Delta$ can be 0 when $|DA|_\Delta$ and $|DB|_\Delta$ are each $\frac{1}{2}$. (Take A to be the Curry sentence K, B to be $\neg K$.)

Returning to the Liar sentence L_0, we now observe that $|DL_0|_\alpha$ is 0, for any $\alpha > 0$; in particular, $|DL_0|_\Delta$ is 0, so $|\neg DL_0|_\Delta$ and $|\neg D\mathsf{True}(\langle L_0 \rangle)|_\Delta$ are 1. So though we can't assert that L_0 isn't true, we *can* assert that it isn't determinately true. Similarly, we can assert that its negation isn't determinately true either; i.e., that the Liar itself isn't determinately false. Similar remarks apply to the Truth-teller T_0.

Once we have noticed the notion of determinateness, it naturally occurs to us to consider an extended Truth-teller sentence T_1, which asserts of itself that it is determinately true; and an extended Liar sentence L_1, which asserts of itself that it is not determinately true. One might guess that T_1 would come out with value $\frac{1}{2}$, on the basis of some sort of principle of indifference; but in fact it gets value 0. Reason: $|DT_1|_0$ is clearly either 0 or $\frac{1}{2}$ by the valuation rules,[14] so $|T_1|_0$ must be the same. This means that $|DT_1|_1$ is 0; so $|T_1|_1$ must be 0. It is evident that at all levels the values of DT_1 and T_1 remain at 0.

In the case of the extended Liar, we require that for each α,

$(*)\ |L_1|_\alpha = 1 - |DL_1|_\alpha$.

How is this possible? (We know that it *is* possible, because D has been defined in the language $\mathcal{L}^+$, and we have given a semantics for $\mathcal{L}^+$ that validates number theory and gives the equality of $\Psi(\mathsf{True}(\langle A \rangle))$ and $\Psi(A)$ within each super-stage.) Clearly there is no α for which $|L_1|_\alpha$ is 0, for then (*) would yield that $|DL_1|_\alpha$ is 1, in violation of Observation (a). Also, when $|L_1|_\alpha$ is $\frac{1}{2}$, $|DL_1|_{\alpha+1}$ is 0 by the evaluation rules, so $|L_1|_\alpha$ is 1 by (*); and when $|L_1|_\alpha$ is 1, $|DL_1|_{\alpha+1}$ is $|L_1|_{\alpha+1}$ by the evaluation rules, so $|L_1|_\alpha$ is $\frac{1}{2}$ by (*). $|DL_1|_0$ is $min\{\frac{1}{2}, |L_1|_0\}$ by the evaluation rules, so by (*) it must be $\frac{1}{2}$; and an easy induction shows that the value at limits is always $\frac{1}{2}$ as well. In short: whenever ζ is 0 or a limit and k is finite (possibly 0), $|L_1|_{\zeta+2k} = \frac{1}{2}$ and $|L_1|_{\zeta+2k+1} = 1$.

In particular: $|L_1|_\Delta = \frac{1}{2}$; so by (*), $|DL_1|_\Delta$ is $\frac{1}{2}$ also. So not only can't we assert $\neg\mathsf{True}(\langle L_1 \rangle)$, we can't assert $\neg D(\mathsf{True}(\langle L_1 \rangle))$ either. That might seem like a defeat. However, we *can* assert that L_1 isn't *determinately* determinately true. For DDL_1 (which I'll abbreviate D^2L_1) takes value 0 at all successor ordinals; this suffices for its taking value 0 at limits as well, i.e. for $\left|\left|\neg D^2 L_1\right|\right|$ and hence $\left|\left|\neg D^2(\mathsf{True}(\langle L_1 \rangle))\right|\right|$ to be 1.

$||\neg D\neg L_1||$ is already 1: we don't need to go to the second level in that case. So the moral about L_1 is: it isn't determinately false, and it isn't *determinately*

[14] In fact it's $\frac{1}{2}$, since we've used the minimal fixed point at each stage.

determinately true.[15] This seems to be fully capture the sense in which it is "defective".

Transfinite iteration: Once we have the operator D, we can iterate it, and get generalized Liar sentences corresponding to each iteration. The finite iterations are clear: D^0A is just A; $D^{n+1}A$ is $D(D^nA)$. At the ω^{th} level, we must get the effect of an infinite conjunction of the D^nA, within the resources of the finite language $\mathcal{L}^+$ in which we're working. Since we have a truth predicate for which $\mathsf{True}(\langle B\rangle)$ is intersubsitutable with B for every B, this is fairly easy. Let $D\text{-}Iterate(y,n,x)$ be an arithmetical formulation of the relation of y being the result of prefixing n occurrences of D to a sentence with Gödel code x. We can then define $D\text{-}Iterate_\omega(y,x)$ ("y is the ω^{th} D-iterate of x") to be (the Gödel code of) the sentence $\forall n\forall y[D\text{-}Iterate(y,n,x) \supset \mathsf{True}(y)]$; we then take $D^\omega A$ to be the (sentence whose Gödel code is) the ω^{th} D-iterate of A. It is routine to check that *for each* α, $|D^\omega A|_\alpha$ is the minimum of $\{|D^nA|_\alpha \mid n<\omega\}$.

And now we can extend further. For each finite $k \geq 0$, $D^{\omega+k}A$ is defined as $D^k(D^\omega A)$, and $D^{\omega+\omega}$ (i.e. $D^{\omega\cdot 2}$) as $D^\omega(D^\omega A)$; clearly we can continue in this way to get $D^\alpha A$ whenever $\alpha < \omega^2$. And we can go to ω^2 too, this time by defining in arithmetic the relation $D^\omega\text{-}Iterate(y,n,x)$ meaning that y is the result of prefixing n occurrences of D^ω to a sentence with Gödel code x, and then defining $D^\omega\text{-}Iterate_\omega(y,x)$ in analogy with the above. It is straightforward to extend the technique to give D^α for any $\alpha < \epsilon_0$, where at each limit we meet the following

> Adequacy Condition: *for each* α, $\left|D^\lambda A\right|_\alpha$ is to be the minimum of $\{|D^\sigma A|_\alpha \mid \sigma<\lambda\}$.

Extension to ϵ_0 or beyond requires more complicated devices, but is possible if the ground language $\mathcal{L}$ contains sufficient resources. However, well-known results on the theory of ordinal notations (see [11] Ch. 11) make clear that we cannot extend beyond the recursive ordinals (a "small" subset of the countable ordinals) on any satisfactory extension procedure; indeed, since a "satisfactory extension procedure" is presumably univalent and recursively related,[16] we can't even get all of the recursive ordinals. (For essentially the same reason, we cannot define a single notion of D^σ for variable σ; the construction of the different D^σ is highly non-uniform, becoming more and more complicated for larger σ.) In what follows I will leave the details of the extension procedure and how far it extends unspecified; whatever the details, the set of σ for which D^σ is defined will be of form $\{\sigma \mid \sigma < \lambda_0\}$, where λ_0 is some recursive limit ordinal.

Generalized Observations: the "Observations" (a)-(c) and (e) of several pages back remain true when 'D' is replaced by 'D^σ', for any $\sigma < \lambda_0$; similarly for (d) if we replace the condition that $\alpha < \beta$ by the condition that $\alpha + \sigma \leq \beta$. (Proof: transfinite induction based on the previous "Observations", using the fact that at each $\lambda < \lambda_0$, $\left|D^\lambda A\right|_\alpha$ is the minimum of $\{|D^\sigma A|_\alpha \mid \sigma<\lambda\}$.)

[15] It's not being determinately false implies, of course, that it is also not *determinately* determinately false.

[16] That is, any ordinal with a notation has a unique canonical one, and there is a mechanical procedure for telling of two canonical notations which stands for the smaller ordinal. If the system of notation failed to meet the first condition, we would really have no business talking about the D^σ for ordinals σ, but rather about the D^ν for various ordinal notations ν.

I will now show that the construction of the D^σ does not terminate: for every $\sigma < \lambda_0$, there are A such that $|D^\sigma A|_\Delta$ is $\frac{1}{2}$ and yet $\left|D^{\sigma+1}A\right|_\Delta$ is 0; hence, for which $\neg D^{\sigma+1}A$ is correctly assertable, while $\neg D^\sigma A$ isn't. $\{D^\sigma \mathsf{True}(x) \mid \sigma < \lambda_0\}$ *is thus a sequence of "stronger and stronger truth predicates"*, with no strongest (since D_0^λ is not defined). (For $\sigma > 0$, we may not want to call them "truth predicates", since only $\mathsf{True}(\langle A \rangle)$ is intersubstitutable with A for all A.) It's important to emphasize that *these are all in the object language: no ascent to stronger and stronger metalanguages is required.*

To see that the D^σ's become successively stronger, it suffices to consider for each $\sigma < \lambda_0$ the corresponding generalized Liar sentence L_σ. L_σ is just $\exists w[\mathsf{SA}(w, \mathbf{k}) \wedge \neg D^\sigma \mathsf{True}(w)]$, where $\mathbf{k}$ is the standard name of $\exists w[\mathsf{SA}(w, v_1) \wedge \neg D^\sigma \mathsf{True}(w)]$; so L_σ is equivalent to $\neg D^\sigma \mathsf{True}(\langle L_\sigma \rangle)$, and so by the properties of the truth predicate that we have demonstrated,

(**) For each α, $|L_\sigma|_\alpha = 1 - |D^\sigma L_\sigma|_\alpha$.

I now show:

Theorem on Transfinite Liar Hierarchy: For each $\sigma < \lambda_0$,

(a) $|D^\sigma L_\sigma|_\Delta = \frac{1}{2}$;

(b) For all $\alpha > 0$, $\left|D^{\sigma+1} L_\sigma\right|_\alpha = 0$;

(c) For all $\alpha > 0$, $|D\neg L_\sigma|_\alpha = 0$ (hence $\left|D^{\sigma+1}\neg L_\sigma\right|_\alpha = 0$).

Proof: (b) By (**) and Generalized Observation (a), $|L_\sigma|_\alpha$ can only be $\frac{1}{2}$ or 1; so by (**) again, $|D^\sigma L_\sigma|_\alpha$ can only be $\frac{1}{2}$ or 0; so for all $\alpha > 0$, $\left|D^{\sigma+1} L_\sigma\right|_\alpha = 0$.

(a) As just noted, $|L_\sigma|_\Delta$ can only be $\frac{1}{2}$ or 1. But if it were 1, then by Generalized Observation (c), $|D^\sigma L_\sigma|_\Delta$ would be 1 also, violating (**). So $|L_\sigma|_\Delta$ is $\frac{1}{2}$, so by (**) $|D^\sigma L_\sigma|_\Delta$ is $\frac{1}{2}$.

(c) By (**), $|\neg L_\sigma|_\alpha = |D^\sigma L_\sigma|_\alpha$. So $|D\neg L_\sigma|_\alpha = \left|D^{\sigma+1} L_\sigma\right|_\alpha$; by (b), that's 0 for $\alpha > 0$. ■

Let's define a sequence of predicates N^σ (for $\sigma < \lambda_0$; 'N' can be read as "nonfactual") by:

$N^\sigma(x)$ iff SENT(x) $\wedge \neg D^\sigma(\mathsf{True}(x)) \wedge \neg D^\sigma(\mathsf{False}(x)))$.

From (a) and (c) of the Generalized Observations, we get that when $\sigma < \rho$, $||N^\rho(\langle A \rangle)|| \geq ||N^\sigma(\langle A \rangle)||$, and that a strict inequality can occur only when the first is 1 and the second $\frac{1}{2}$. Our observations on the Liar sentence are that $||N^\sigma(\langle L_\sigma \rangle)||$ is $\frac{1}{2}$ and $\left|\left|N^{\sigma+1}(\langle L_\sigma \rangle)\right|\right|$ is 1, so we have that as σ increases these non-factuality predicates get ever more inclusive in their positive extension for as long as they are defined (i.e. up to λ_0), and that each of the generalized Liar sentences is eventually in the positive extensions. And these predicates are all present in the object language, definable via set theory from the truth predicate, using the '$\rightarrow$'.

Illustrative details: That completes the main point of this section, but it is instructive to see a few illustrations of what values various paradoxical sentences assume at various stages.

First, what are the values of the generalized Liar sentences L_σ? (The answer to this will prove something that may be important: that any acceptable Δ is at least λ_0.) The values are as follows: for any $\sigma < \lambda_0$ and any α,

$$|L_\sigma|_\alpha = \begin{cases} \frac{1}{2} & \text{whenever } \alpha \text{ is of form } (1+\sigma)\cdot\tau \\ 1 & \text{otherwise.} \end{cases}$$

We've already seen that this satisfies (**) in the $\sigma = 0$ case (the ordinary Liar, which has value $\frac{1}{2}$ for all α). For $\sigma \geq 1$: let $rm(\alpha)$ be the remainder of α on left-division by $1+\sigma$; that is, the unique $\rho \leq \sigma$ such that α is of form $[(1+\sigma)\cdot\tau]+\rho$. We can show, by induction on β, that

(***) for any β in $[0, \sigma]$ and any α,

$$\left|D^\beta L_\sigma\right|_\alpha = \begin{cases} \frac{1}{2} & \text{when } rm(\alpha) = 0; \\ 0 & \text{when } 1 \leq rm(\alpha) < 1+\beta; \\ 1 & \text{when } 1+\beta \leq rm(\alpha) < 1+\sigma. \end{cases}$$

(Proof in footnote.)[17] Once this is established, then taking β to be σ we get

$$|D^\sigma L_\sigma|_\alpha = \begin{cases} \frac{1}{2} & \text{whenever } \alpha \text{ is of form } (1+\sigma)\cdot\tau \\ 0 & \text{otherwise;} \end{cases}$$

that is, $1 - |L_\sigma|_\alpha$, as required.

Since for each $\sigma < \lambda_0$, L_σ doesn't reach its final value until stage $\sigma+1$, any acceptable ordinal must be at least as big as $\sigma+1$ for each $\sigma+1 < \lambda_0$; that is, it must be at least λ_0.

The two-dimensional Curry Hierarchy. I now consider the Curry paradox, and generalizations of it. The standard Curry sentence K (which I will also write as $K_{0,1}$) asserts of itself that if it is true then $\perp$, where $\perp$ is some absurdity. So for any α, $|K|_\alpha = |K \rightarrow \perp|_\alpha$. It is easy to see that this requires that $|K|_\alpha$ is $\frac{1}{2}$ when α is 0 or a limit, 0 when α is odd, 1 when it is an even successor. $|K \rightarrow \perp|_\alpha$ comes out having the same values, as required. $|DK|_\alpha$ is 0 except at 0 and limits, where it is $\frac{1}{2}$. (Note that it is not continuous at limits until ω^2.) D^2K is 0 after stage 0.

There are several ways in which we might generalize this. One is analogous to what we did with the Liar: we consider a sentence that asserts of itself that

[17] Establishing (***) by induction: (I) For $\beta = 0$, (***) reduces to the description of $|L_\sigma|_\alpha$. (II) *Successors.* Suppose it holds for $\beta = \gamma$, where $\gamma < \sigma$. To establish it for $\beta = \gamma+1$, we must deal with three cases. *Case 1:* $rm(\alpha) = 0$. The induction hypothesis says that $|D^\gamma L_\sigma|_\alpha$ is $\frac{1}{2}$; so to show that $\left|D^{\gamma+1} L_\sigma\right|_\alpha$ is $\frac{1}{2}$, we merely need to show that for any $\delta < \alpha$, there are ρ in $[\delta, \alpha)$ for which $|D^\gamma L_\sigma|_\rho$ is 1. This is trivial for $\alpha = 0$. Otherwise, write δ as $(1+\sigma)\cdot\mu + \upsilon$, where $\upsilon < 1+\sigma$. Since α is $(1+\sigma)\cdot\tau$, $\mu < \tau$; and since $\gamma < \sigma$, $1+\gamma < 1+\sigma$, so $(1+\sigma)\cdot\mu + 1 + \gamma < \alpha$. But the induction hypothesis also says that whenever $\rho \geq 1+\gamma$, $|D^\gamma L_\sigma|_{(1+\sigma)\cdot\mu+\rho}$ is 1; so letting ρ be $max\{\upsilon, 1+\gamma\}$, we have the desired ρ in $[\delta, \alpha)$ for which $|D^\gamma L_\sigma|_\rho$ is 1. *Case 2:* $1 \leq rm(\alpha) \leq \gamma+1$. Then $\left|D^{\gamma+1} L_\sigma\right|_\alpha$ is 0: for all values of $rm(\alpha)$ except for $\gamma+1$, this follows from the fact that $|D^\gamma L_\sigma|_\alpha$ is 0; and when $rm(\alpha)$ is $\gamma+1$, $\left|D^{\gamma+1} L_\sigma\right|_\alpha$ is 0 since then α is $\delta+1$ for a δ for which $rm(\delta)$ is γ and hence $|D^\gamma L_\sigma|_\delta$ is 0. *Case 3:* $\gamma+1 < rm(\alpha) \leq \sigma$. Then $\left|D^{\gamma+1} L_\sigma\right|_\alpha$ is 1, since for all δ in $[\gamma+1, \alpha)$ and also for $\delta = \alpha$, $|D^\gamma L_\sigma|_\delta$ is 1. (III) *Limits.* Suppose (***) holds for all $\beta < \lambda$, where λ is a limit $\leq\sigma$. By the adequacy condition on treatment of limits, $\left|D^\lambda L_\sigma\right|_\alpha$ is $min\{\left|D^\beta L_\sigma\right|_\alpha \mid \beta < \lambda\}$; by induction hypothesis that's $\frac{1}{2}$ iff $rm(\alpha)$ is 0, and it's 0 iff for some $\beta < \lambda$, $1 \leq rm(\alpha) \leq 1+\beta$, that is, if $1 \leq rm(\alpha) < 1+\lambda (= \lambda)$.

if it is D^σ-true then $\bot$. Call that sentence $K_{\sigma,1}$. Another way to generalize is to iterate the conditional. For each k, define $A \rightarrow^k B$ as follows: $A \rightarrow^0 B$ is B; $A \rightarrow^{k+1} B$ is $A \rightarrow (A \rightarrow^k B)$. We can also extend this into a portion of the transfinite, taking $A \rightarrow^\lambda B$ to be the "infinite disjunction" of the $A \rightarrow^\rho B$ for $\rho < \lambda$, using the same technique for defining infinite disjunctions as before.[18] (We can either include or exclude the $\rho = 0$ case in the disjunction; it won't matter in the case of the Curry sentences, where B is $\bot$.) A further generalization of the Curry sentence, then, is a sentence $K_{\sigma,\rho}$ that asserts $D^\sigma \mathsf{True}(\langle K_{\sigma,\rho}\rangle) \rightarrow^\rho \bot$. In the case where $\rho = 0$ there is no air of paradox: each $K_{\sigma,\rho}$ is just $\bot$. When $\rho > 0$, these all present distinct paradoxes, tending to become "more paradoxical" as σ and ρ increase. Even sticking to the case where $\sigma = 0$, there is no solution (consistent with the intersubstitutivity requirements on truth) for $\rho = 1$ within classical logic; there is none for $\rho = 2$ within Łukasiewicz 3-valued logic; and when ρ is infinite there is none within Łukasiewicz continuum-valued logic.[19] But again, the result of Section 2 guarantee that these (and the ones with $\sigma > 0$ as well) are all consistently evaluable in the present semantics.

I leave to the reader the full investigation of what values the various $K_{\sigma,\rho}$ sentences take on at various stages, and what the various values are of $D^\sigma K_{\sigma,\rho} \rightarrow^\alpha \bot$ and of $D^\beta K_{\sigma,\rho}$ and $D^\beta \neg K_{\sigma,\rho}$. But the following are easy to check:

$$(1) \qquad |K_{0,k}|_\alpha = \begin{cases} \frac{1}{2} & \text{if } \alpha \text{ is 0 or a limit} \\ 0 & \text{if } \alpha + 1 \text{ is a multiple of } k+1 \text{ but not a limit;} \\ & \text{that is, if } \alpha \text{ is of form } k + n(k+1) \text{ or } \lambda + k + n(k+1) \\ 1 & \text{otherwise} \end{cases}$$

(Explanation: for each finite $j \in \{1, ..., k\}$, we can show that $\left|K_{0,k} \rightarrow^j \bot\right|_\alpha$ is $\frac{1}{2}$ if α is 0 or a limit; 0 if for some $m \leq k - j$, $\alpha + 1 + m$ is a multiple of $k + 1$ but not a limit; 1 otherwise. Taking $j = k$, we get that $\left|K_{0,k} \rightarrow^k \bot\right|_\alpha$ is $|K_{0,k}|_\alpha$, as desired.)

So $\left|D^k K_{0,k}\right|_\alpha$ and $|D\neg K_{0,k}|_\alpha$ are $\frac{1}{2}$ at 0 and limits, 0 everywhere else; so $\left|D^{k+1} K_{0,k}\right|_\alpha$ and $\left|D^2 \neg K_{0,k}\right|_\alpha$ are 0 whenever $\alpha > 0$. Thus $N^{k+1}(\langle K_{0,k}\rangle)$, for finite k other than 0.

$$(2) \qquad |K_{0,\omega}|_\alpha = \begin{cases} \frac{1}{2} & \text{if } \alpha \text{ is 0 or a limit that is divisible by } \omega^2 \\ 0 & \text{if } \alpha \text{ is any other limit} \\ 1 & \text{otherwise} \end{cases}$$

(Explanation: for each finite j, $\left|K_{0,\omega} \rightarrow^j \bot\right|_\alpha$ has the same value at limits as $|K_{0,\omega}|_\alpha$; if $k > 0$ then its value at level k, or $\lambda + k$ when its value at λ is $\frac{1}{2}$, is 1 if $k < j$, 0 if $k \geq j$; and its value at $\lambda + k$, when its value at λ is 0, is 1 if $k \leq j$, 0 if $k > j$. $|K_{0,\omega} \rightarrow^\omega \bot|_\alpha$ is in effect the infinite disjunction of the $|K_{0,\omega} \rightarrow^\omega \bot|_\alpha$, which is $|K_{0,\omega}|_\alpha$ as desired.)

So $|D^\omega K_{0,\omega}|_\alpha$ and $|D\neg K_{0,\omega}|_\alpha$ are 0 except at 0 and limits divisible by ω^2, where they are $\frac{1}{2}$; so $\left|D^{\omega+1} K_{0,\omega}\right|_\alpha$ and $\left|D^2 \neg K_{0,\omega}\right|_\alpha$ are 0 for all $\alpha > 0$, so $N^{\omega+1}(\langle K_{0,\omega}\rangle)$.

[18] This transfinite extension isn't very natural in the general case: we don't in general have $(A \rightarrow^k B) \rightarrow (A \rightarrow^{k+1} B)$, so why should we go to disjunction in the limit? However, it is easily proved by induction that we do have $(A \rightarrow^k B) \rightarrow (A \rightarrow^{k+1} B)$ in the case where B is $\bot$, so the use in the context of the Curry paradox is quite natural.

[19] For the latter result, see [10] or [7].

In addition to the two-dimensional hierarchy of Curry sentences, we could also consider modified Curry sentences where instead of the consequent $0 = 1$ we use another unassertable sentence: for instance, something in the hierarchy of Liar sentences, or another Curry sentence, or whatever. But enough's enough.

6 Ultimate revenge?

It is widely thought that any proposed solution to the Liar paradox faces a "revenge problem": a new paradox, analogous to the Liar, that remains unresolved. The general method of constructing such new paradoxes, for solutions to the old paradoxes that involve a logic based on an extension of Kleene semantics, is to argue that we ought to be able to include in the object language an operator $D^\star$, where $D^\star A$ (or $D^\star\mathsf{True}(\langle A\rangle)$) means '$A$ has value 1'. We then argue, either on semantic or on inferential grounds, that the inclusion of such an operator within the language would lead to a new "hyper-paradox": a sentence for which the truth schema is not satisfiable.

Semantically, the argument is that $D^\star A$ should have value 1 when A has value 1, and should have value 0 when A doesn't have that value; and that excluded middle should hold for attributions of semantic value (since such attributions are made in a classical metalanguage), so that $D^\star A$ always has value 1 or 0. (We also assume that no sentence can have more than one value.) But if $D^\star$ were included in the language, there would be a sentence $L_\star$ that asserts $\neg D^\star\mathsf{True}(\langle L_\star\rangle)$, and hence would be equivalent to $\neg D^\star L_\star$ if the naive truth theory holds. But then the value of $D^\star L_\star$ would be 1 iff it's 0 and 0 iff it's 1; and it's either 1 or 0 and not both, so (using disjunction elimination) this yields a contradiction. So we can't consistently assign a semantic value to $D^\star L_\star$ without violating the truth schema.

Inferentially, the claim is that $D^\star$ should be such that both the inference from A to $D^\star A$ and its converse are valid, and in addition $D^\star A \vee \neg D^\star A$ should be valid (since it is a correct principle of the classical metalanguage). But again, $L_\star$ causes a problem. For from $D^\star L_\star$ we can infer both $L_\star$ and $\neg L_\star$, hence we can infer anything; similarly, we can infer anything from $\neg D^\star L_\star$, since $\neg D^\star L_\star$ implies $L_\star$ which implies $D^\star L_\star$. But then (by disjunction elimination) we can infer anything from $D^\star L_\star \vee \neg D^\star L_\star$; and we've assumed that instance of excluded middle valid, so our principles are hopelessly inconsistent as applied to sentences containing any such $D^\star$.

One possible conclusion from this is that our principles are of only limited validity: the construction of Section 2 shows that they apply unproblematically to restricted object languages, but the claim is that when we try to incorporate the metalanguage within the object language we need to restrict our rules and modify our semantics. Such a conclusion would be disappointing. I will argue that there is no basis for such a conclusion, and that the lessons of the attempt to produce a hyper-paradox are very different.

Before arguing this, I pause to note that we have defined within the language a whole class of operators D^σ that have some of the features of the $D^\star$ above. Semantically, for each σ, $D^\sigma A$ always has value 1 when and only when

A does; but it need not have value 0 when A has a value other than 1. Inferentially, each $D^\sigma A$ is interderivable with A; but excluded middle is not valid for any such D^σ.

Of course, it is no surprise that we can't produce in the language an operator with the features required for a hyper-paradox, for the construction in Section 2 shows that no paradox is actually producible in the language (that is, it shows that the naive truth theory is satisfiable in the language). What the proponent of hyper-paradox claims is (to put it vaguely) that it is only because of an expressive limitation of the language that paradox has been avoided. Is there any basis for this charge?

One unpromising approach to substantiating the charge would be to focus on the fact that although we can define each of the D^σ within the language, we cannot define their "infinite conjunction" D^{λ_0}. Let's grant, for the sake of argument, that the "infinite conjunction" of these is intelligible, so that the inexpressibility of such an infinite conjunction in the language is indeed a genuine expressive limitation of the language.[20] If we grant this, then we can imagine a more powerful language that includes a "superdeterminately" operator D^{λ_0}, together with the noneffective set of principles that is required to ensure that it works like an infinite conjunction of the previous D^σs; this more powerful language contains a "super-Liar" sentence that says of itself that it is not superdeterminately true. But in such a more powerful language, we should be able to reason analogously to the way we've reasoned already, to show that such a sentence is not paradoxical: the claim that it is not superdeterminately true would be neither assertable nor deniable, but we could assert that it is not *determinately* super-determinately true, i.e. not D^{λ_0+1}-true. In other words, there's no reason to think that overcoming expressive limitations in this way would change anything important, so there's no reason to think that it is only through expressive limitations that paradox has been avoided.

The basic defect of the approach to hyper-paradox in the previous paragraph (and more complicated approaches that iterate the approach in the previous paragraph) is that there is no reason whatever to think that the expansions of the language they envision will meet the conditions required for a hyper-paradox: a paradox not soluble along just the same lines as the paradoxes that are handled within the language. In particular, there is no reason to suppose that the value of $D^{\lambda_0} A$ should be 0 whenever the value of A is $\frac{1}{2}$, which undermines the semantic argument for hyper-paradox; and there is no reason to suppose that excluded middle should hold for sentences of form $D^{\lambda_0} A$, undermining the inferential argument as well. If excluded middle did have to hold for such sentences (and in addition we could infer from $D^{\lambda_0} A$ to A and conversely, and reason by disjunction-elimination), then the introduction of D^{λ_0} would give rise to the inferential form of the hyper-paradox, and it would indeed be the case that only expressive limitations were preventing a violation of the truth schema. But there's no reason for thinking that excluded middle should hold for such sentences.

[20] There might be grounds for doubting what I'm granting: the non-uniformity of how the different D^σ are constructed, which precludes introducing σ as a quantifiable variable, might give some reason to doubt that we really grasp the infinite conjunction. But I wouldn't want to press that point.

But what I've just said simply shows that it was a mistake to try to use such a D^{λ_0} for the hyper-paradox. The hyper-paradox as originally sketched was in terms of an operator $D^\star$ defined directly in terms of semantic value. We have defined semantic values for our sentences in a classical set-theoretic metalanguage, a metalanguage for which, we are assuming, excluded middle holds. Doesn't this show that we ought to be able to expand the object language so as to incorporate this classical metalanguage, enabling us to define an operator or predicate corresponding to 'has semantic value 1', for which a hyper-paradox is bound to arise?

I think this thought is an illusion. The first point to notice is that the construction of Section 2 was all done within ZFC (Zermelo-Frankel set theory with choice), and I was careful to insist that the object language $\mathcal{L}$ from which I began (before I added 'True' and '$\rightarrow$') was any classical language that included arithmetic and the general theory of finite sequences.[21] In particular, I allowed the language $\mathcal{L}$ to be the language of ZFC (or any classical expansion of that, e.g. to include the language of physics). If that is the $\mathcal{L}$ from which we start, and if the arithmetically standard model M from which we start is definable within $\mathcal{L}$ (as will be the case for the most natural choices of M, e.g. for all inner models), then there is no need to use a *broader* classical metalanguage to do the semantics; we can use $\mathcal{L}$ itself. Thus the classical metalanguage is not an expansion of the object language, it is in fact a *sublanguage* of the nonclassical object language $\mathcal{L}^+$, it is the part of the object language that doesn't include 'True' and '$\rightarrow$'. (The sublanguage is classical in that excluded middle is explicitly postulated to hold in it.)

If that's so, the question naturally arises: how have we avoided paradox? If the metalanguage is included in the object language, then we can construct within the object language a sentence that says of itself that it does not have semantic value 1. Since this metalanguage is a classical part of the object language, we are assuming excluded middle for it; so haven't the preconditions for paradox been met?

No. The reason why they haven't becomes clear when we reflect on Tarski's theorem about the impossibility of *defining* within a classical language a predicate that satisfies the requirements of a truth predicate for that classical language. The theorem shows that the predicate 'has semantic value 1' that we've defined within set theory (when the starting model M is so definable) doesn't give a reasonable notion of truth *even for the set theoretic language itself*, let alone for the expanded nonclassical language that results from adding 'True' and '$\rightarrow$' to it. (Nor does it give a reasonable notion of *determinate* truth for the set theoretic language: indeed, there is no obvious reason to distinguish determinate truth from plain truth, for sentences in a classical language, which is what we are assuming the set-theoretic sublanguage to be.) More fully, Tarski's argument shows that on any set-theoretic definition of 'has semantic value 1', there will always be a sentence S in the set-theoretic language itself (i.e., one not containing 'True' or '$\rightarrow$') for which we can prove

[21] By a "classical language" I mean simply one for which classical reasoning is appropriate. Strictly speaking, the term 'classical' could be dropped from the sentence in the text: but since we considered only classical models for the langage, the construction in Section 2 would be unnatural if the starting language were not classical in the sense just given.

$[(\langle S\rangle$ has semantic value 1) $\wedge\neg S] \vee [S \wedge \neg(\langle S\rangle$ has semantic value 1)].

It is clear from Tarski's theorem, then, that no notion of semantic value that is *defined within the classical sublanguage* can possibly coincide with the intuitive notion of truth (or determinate truth), even restricted to that classical sublanguage. This doesn't show that we can't introduce a notion of truth or determinate truth that works when restricted to the classical sublanguage: we can certainly introduce such notions axiomatically (and I've in fact argued that we can introduce reasonable notions of determinate truth definitionally, if we start from an undefined notion of plain truth). The problem is simply a limitation in *definitions* of truth or determinate truth from a basis that includes no such notion.

Tarski did, of course, show how to literally define a notion of *truth in a model*, where a model is an object within the universe of sets: the domain of any model M is a set, and since (according to the set theory presupposed in Tarski's definition) there is no set of all sets, no model can include everything in its domain.[22] What Kripke showed was how to extend this definition of truth-in-a-model in a natural way to a larger language, one containing 'True'; what I showed in Section 2 was how to further extend it to a still larger language that contains a new conditional. But it is important to be clear that what is being defined is not 'true', but 'true in the starting model M': that's why I made a point of not calling the defined notion 'true', and using instead the phrase 'has semantic value 1'. Semantic value, in the sense I've defined it (and in the sense in which Tarski or Kripke define such notions), depends on the starting model M. That's why I (and Kripke before me) considered languages in which 'True' was an additional undefined term.[23]

We can do a lot to ensure that the model M we start with is an extremely "natural" one: for instance, if our basic theory is an expansion of ZFC that postulates inaccessible cardinals (or hyper-inaccessible cardinals, etc.), then we can take as our starting model M the set of all sets of accessible (or hyper-accessible, etc.) rank (i.e. of rank less than the initial ordinal of the first such cardinal), with the standard membership relation. If we do this, and let 'has semantic value 1' as applied to set-theoretic sentences just mean 'is true in M' in the Tarskian sense, then for any sentence *A all of whose quantifiers are restricted by the condition 'has accessible rank'*, we can prove

$\langle A\rangle$ has semantic value 1 if and only if A.

[22] One could give a Tarskian definition of truth in a model in a theory that allows for proper classes, but (assuming we keep excluded middle for the theory of proper classes) this wouldn't alter much: while we could define true in M where M is a proper class model that includes all sets, there is still no proper class model that includes all classes, so what we've defined still won't be truth, if the proper class theory in which we give the definitions is true. So we may as well just stick to sets.

[23] Kripke's own presentation is a bit misleading on this score: indeed, he insists that he is offering an explicit definition of a truth predicate, and criticizes those who offer less. But in fact what he explicitly defines is only truth in a model; that is all that one could hope to explicitly define, given Tarski's undefinability theorem. (The nonclassicality of the language doesn't evade Tarski's theorem, since the language contains a classical part with the strength required for the theorem to apply.)

In other words, *for sentences with quantifiers so restricted*, semantic value 1 (truth in M) will coincide with real truth for sentences in the set-theoretic language. Even so, it does not correspond with truth (or determinate truth, or any such thing) across the board: for the theory implies the existence of inaccessibles, but also implies that the sentence S asserting the existence of inaccessibles gets semantic value 0.[24]

Once we see that there is an inevitable gap between having semantic value 1 in any *defined* sense and truth or determinate truth, we see that the hyper-paradox (in the form so far discussed) dissolves. The classical set-theoretic metalanguage, recall, is a proper part of the object language of primary interest (which includes not only full set theory but also 'True' and '$\rightarrow$'). We're assuming that excluded middle holds within this restricted metalanguage. Within this restricted metalanguage, we can also prove that every sentence of the object language, and *a fortiori* of the metalanguage, has exactly one of the semantic values 0, $\frac{1}{2}$ and 1; in particular, this is so of the "hyper-liar" sentence H that asserts of itself that it does not have semantic value 1. Letting SV_M represent semantic value (relative to the ground model M), the obvious "paradoxical reasoning" proves the following disjunction:

$$\text{Either } SV_M(\langle H\rangle) = 1 \text{ and } \neg H, \text{ or } SV_M(\langle H\rangle) \neq 1 \text{ and } H.$$

Or employing the truth schema (which is completely uncontroversial in this instance since H is simply a sentence of set theory)

> (*) Either $SV_M(\langle H\rangle) = 1$ and $\langle H\rangle$ isn't true, or $SV_M(\langle H\rangle) \neq 1$ and $\langle H\rangle$ is true.

While the conclusion may at first seem surprising, it is really just an instance of what we've seen is inevitable when we try to use a defined predicate like 'has semantic value 1' as a surrogate for truth: we will always get extensional failures. Indeed, the reason for the failure is essentially the same as in the illustration involving inaccessible cardinals: in defining 'has semantic value 1 (relative to M)', we must inevitably employ unrestricted quantifiers, quantifiers that range over sets not included within the domain of M. So the hyper-liar H is a sentence that essentially involves unrestricted quantifiers, and so there's no reason why the defined surrogate for truth should be expected to correspond to real truth in that case. (Which of the two disjuncts of (*) holds presumably

[24] A similar point arises for ground models not definable in $\mathcal{L}$, but definable only in broader classical sublanguages of $\mathcal{L}^+$. For instance, if we add to the language of ZFC a predicate 'is a true sentence of ZFC', it is possible to define in this extended language ZFC* a (countable) arithmetically standard model (ω-model) of ZFC that "reflects the real universe" with respect to sentences in the language of ZFC; that is, the sentences of the language of ZFC that are true in M are precisely those that are genuinely true. (The definition incorporates a downward Löwenheim-Skolem construction on the full universe; the resulting model is thus quite "unnatural".) If we perform the construction of semantic value from such a starting model, there will be no sentence *of ZFC* for which there is a gap between having semantic value 1 and genuine truth; but Tarski's Theorem still shows that there is a sentence *of ZFC** for which there is such a gap, which is again enough to show that semantic value 1 in the defined sense can't quite coincide with truth. And again, because of the classical nature of such a restricted truth predicate, truth and determinate truth should coincide, so semantic value 1 can't coincide with determinate truth either. (Thanks to Robert Black for a question that inspired this footnote, and to John Burgess and Stewart Shapiro for discussion of some points related to it.)

depends on the vagaries of the starting model M; but whichever of the two disjuncts holds for H, the other disjunct holds with $\neg H$ substituted everywhere for H.)

I do not mean to suggest that the notion of having semantic value 1 in the sense defined has nothing to do with truth or determinate truth. On the contrary, it serves as a good model of these notions (in an informal sense of model): just as our starting model M (in the technical sense) is (in an informal sense) a slightly inaccurate model of the full universe of sets, so truth in M (i.e. having semantic value 1, relative to M) is a slightly inaccurate model of genuine truth. Because it is a model of it *in a classical metalanguage*, it is inevitably a feature of the model that all questions have determinate answers: for any sentence, either it has semantic value 1 relative to M or it doesn't. But we have independent reason to know that the model cannot be taken seriously in all respects; and the fact that attributions of semantic value satisfy excluded middle is one of the respects in which the model can not be taken seriously. (Having semantic value 1 is in fact a bit better as a model of determinate truth than of truth, and of determinate determinate truth than determinate truth, and so forth: for when $\rho > \sigma$ there are more A for which $D^\rho A \vee \neg D^\rho A$ holds than for which $D^\sigma A \vee \neg D^\sigma A$ holds. But we know from the above discussion of Tarski's theorem that having semantic value 1 relative to M can't possibly correspond to any reasonable notion of determinate truth, even for sentences of pure set theory where any reasonable notion of determinateness is presumably redundant.)

I have been arguing (following Tarski) against the possibility of *defining* notions of a truth-theoretic sort from a basis that excludes such notions: any such attempt yields at best an approximation to the notions we are really aiming at. But might we introduce slightly different notions of having semantic values 1, $\frac{1}{2}$, and 0 purely by axioms, where in contrast to the notions defined before, these are conceived as not dependent on a starting model but as "related to the real universe as the defined notions were related to the ground model M"? Yes, I think it can be easily done. For instance, we could take as primitive the predicate $RSV_1(x)$ of being a sentence of $\mathcal{L}^+$ with "real" (not model-dependent) semantic value 1, define $RSV_0(x)$ as $RSV_1(neg(x))$ and $RSV_{\frac{1}{2}}(x)$ as $SENT(x) \wedge \neg RSV_1(x) \wedge \neg RSV_0(x)$, and introduce the following axioms and rules:

$A \models RSV_1(\langle A \rangle)$, when A is a sentence of $\mathcal{L}^+$

$\models RSV_1(\langle A \rangle) \supset D^\sigma A$, when A is a sentence of $\mathcal{L}^+$ (a separate such axiom for each $\sigma < \lambda_0$).

No finite subset of these axioms can lead to paradox, since we can always then interpret RSV_1 as the highest D^σ in the set, so there will certainly be no paradox derivable in the inferential system given.

I'm not sure that there is a great deal of value in adding an axiomatic notion of "real semantic value" to the language: I doubt that we really understand such a notion, and also doubt that developing an understanding of one would serve much purpose. But for what it's worth, it might be possible to consistently extend the system just described, in a number of ways. One way to extend it would be to include laws governing the interaction between the

new predicate RSV_1 and the $\rightarrow$ operator. Such laws would be very important to understanding the significance of the new RSV_1 predicate: for instance, they would determine such things as whether $D(RSV_1(x))$ is stronger than $RSV_1(x)$ and whether $D(\neg RSV_1(x))$ is stronger than $\neg RSV_1(x)$. Another way to extend it—though one that reinforces doubts that we understand what we're doing—would be to allow the displayed rules of the previous paragraph to apply even when A was a sentence of the enlarged language that includes RSV_1 (in which case we could modify the definition of $RSV_{\frac{1}{2}}(x)$ by replacing '$SENT$', i.e. 'sentence of $\mathcal{L}^+$', by something meaning 'sentence of the enlarged language').[25] Doing this would lead to inferential hyper-paradox if we also assumed excluded middle for all instances of the predicate $RSV_1(x)$ or $RSV_{\frac{1}{2}}(x)$ (and continued to assume disjunction-elimination in the expanded language). But now that we have given up on the project of defining $RSV_1(x)$ in the classical set-theoretic language (or the language of it together with a purely classical truth predicate), there's no reason to suppose that excluded middle holds for the predicate $RSV_1(x)$ or $RSV_{\frac{1}{2}}(x)$. And the fact that excluded middle would lead to paradox is about as strong reason against assuming it as could be desired.

My skepticism about the existence of a deep philosophical role for a *definition* of the notion semantic value might seem to sit ill with the fact that such definitions played a major role in this paper: notably, in Section 2. In fact, though, there is no conflict: the model-relative definition of semantic value is precisely what is needed for the main goal I introduced it for, which was to provide a strong sort of relative consistency proof for the naive theory of truth in the logic I've set out. What I've done in effect is to show in set theory that for any theory T that includes the arithmetic required to develop formal syntax, and any classical model M of T that is standard with respect to that arithmetic, one can define a new non-classical model for the logic LCC which satisfies naive truth theory. Moreover, the domain of the new model is precisely the same as that of the old, and all of the predicates of T have the same extension in the new model as on the old; among other things, this implies that the new model is standard with respect to arithmetic too. So if T is consistent with respect to ω-logic (the logic of arithmetically standard models), so is the theory $T^\star$ in LCC whose axioms are those of T plus the axioms of naive truth theory.[26] Since this holds for *any* theory T with an arithmetically standard model—including a theory that includes claims about the physical world—this shows not just the consistency of naive truth theory in LCC, but its "conservativeness", in one important sense of that phrase.[27]

[25] Whether the result of so doing would be significantly different from adding the D^{λ_0} predicate contemplated earlier would depend on the laws of interaction between RSV_1 and '$\rightarrow$'.

[26] T itself can be conceived either as a classical theory (supplemented by the ω-rule) or as a theory in LCC (supplemented by the ω-rule); a classical theory is just a special case of a theory in LCC, it is a theory in LCC in which excluded middle is assumed for all instances of all the basic predicates.

[27] The reason for the qualification: many people think that to call naive truth theory conservative should require that whenever T is consistent, then $T^{\star\star}$ should be consistent, where $T^{\star\star}$ has as its axioms those of $T^\star$ *plus all instances of any schemas in* T *that contain the new term* 'True'. Naive truth theory is certainly *not* conservative in the latter sense. Indeed, the extension of the schemas of number theory or set theory to instances involving 'true' is slightly delicate: if paradox is to be avoided, the schematic axiom generally needs to be replaced by a schematic rule, whose instances imply the corresponding instances of the axiom in contexts where excluded middle is assumed: see [5] for some discussion of this. Even with schemas in the rule form, the "conservativeness" of naive truth theory does not in itself guarantee that we can extend the schemas of ZFC to predicates

Given that our goal is this sort of "conservativeness" proof, the point of the definition of semantic value is simply to enable us to construct the non-classical model from the classical; and so the dependence of the notion of semantic value on the starting model is no cause of concern. The construction of the new model uses standard set theory (Zermelo-Fraenkel with choice), so of course it only yields the consistency (and "conservativeness") of naive truth theory relative to that set theory; but that seems a pretty solid consistency proof nonetheless.

I don't mean to suggest that the only desideratum on an acceptable logic for the paradoxes is that there be a "conservativeness" proof for naive truth theory in the logic. (For instance, the logic of [5] would pass this test, but its conditional seems inadequate in various ways, e.g. in failing to reduce to the material conditional in contexts where excluded middle is assumed. The logic of [2] passes the test too, but by my lights it is excessively weak, e.g. in being a relevance logic.) And I don't deny that in selecting among logics in which naive truth theory is "conservative", a definition of semantic value appropriate to one such logic could serve a heuristic role in motivating that logic over the others. But the main desideratum in making such a selection isn't the notion of semantic value, but the ease of working with the logical laws that are validated.

I make no claim that LCC is optimal in this respect: I wouldn't be at all surprised if there were alternative approaches to the conditional, based on alternative semantics, that were more attractive overall. (E.g., they might include "conditional strengthenings" of some of the B and D axioms, and this gain in strength might be more appealing than any losses that were required to compensate for the gain.) I do think, though, that LCC is a significant improvement over any of the other attempts to preserve naive truth theory currently in the literature.[28]

References

[1] George Boolos. *The Logic of Provability*. Cambridge University Press, Cambridge, 1993.

[2] 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.

[3] Hartry Field. On conservativeness and incompleteness. *Journal of Philosophy*, 81:239–60, 1985.

[4] Hartry Field. Deflating the conservativeness argument. *Journal of Philosophy*, 96:533–40, 1999.

that contain the term 'True' without generating an inconsistency (with respect to ω-logic); though the "picture" that makes the consistency of ZFC (with respect to ω-logic) plausible also makes the consistency of this extension (with respect to ω-logic) plausible. (For further discussion of the notion of conservativeness in connection with schematic theories, see [4], which is a response to [13] and implicitly to [8]; and [3], a response to [12].)

[28] Besides those mentioned in n. 24, I'd like to thank Thomas Hofweber, Kit Fine and Joshua Schechter for suggesting improvements.

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

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

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

[8] Jeffrey Ketland. Deflation and tarski's paradox. *Mind*, 108:69–94, 1999.

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

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

[11] Hartley Rogers. *Theory of Recursive Functions and Effective Computability.* McGraw-Hill, New York, 1967.

[12] Stewart Shapiro. Conservativeness and incompleteness. *Journal of Philosophy*, 80:521–31, 1983.

[13] Stewart Shapiro. Proof and truth: Through thick and thin. *Journal of Philosophy*, 95:493–521, 1998.