DOES TRUTH BEHAVE LIKE A CLASSICAL CONCEPT WHEN THERE IS NO VICIOUS REFERENCE?

Philip Kremer, Department of Philosophy, University of Toronto

Abstract. In *The Revision Theory of Truth* (MIT Press, 1993), Gupta and Belnap claim that one advantage of their particular approach to truth is "its consequence that truth behaves like an ordinary classical concept under certain conditions—conditions that can roughly be characterized as those in which there is no vicious reference in the language." Aiming to clarify this remark, they define *Thomason* models, nonpathological models in which truth behaves like a classical concept. They investigate conditions under which a model is Thomason, arguing a model is Thomason when there is no vicious reference in it. In the current paper, we extend their investigation, considering notions of nonpathologicality and senses of "no vicious reference" generated both by revision theories of truth and by fixed point theories of truth. We show that some of the fixed point theories have an advantage analogous to that which Gupta and Belnap claim for their approach, and that at least one revision theory does not. This calls into question the claim that the revision theories have a *distinctive* advantage in this regard.

§1. Introduction. When truth-theoretic paradoxes are generated, two factors seem to be at play: the behaviour that truth intuitively has; and the facts about which singular terms refer to which sentences, and so on. For example, paradoxicality might be partially attributed to the contingent fact that the singular term, "the italicized sentence on page one", refers to the sentence,

The italicized sentence on page one is not true.

Factors of this second kind might be represented by a *ground* model: an interpretation of all the names, function symbols, and predicates in the potentially self-referential language under study, with the exception of the predicate "*x* is true". Formally, suppose that *L* is an uninterpreted first order language. M = ⟨D, I⟩ is a *classical model* for *L* if D is a nonempty set; and I is a function assigning to each name of *L* a member of D, to each n-place function symbol of *L* a function from D^n to D, and to each nonlogical n-place predicate of *L* a function from D^n to {**t**, **f**}. Suppose furthermore that L^+ is obtained by adding a new one-place predicate ***T*** to *L*, and that *L* has a quote name '*A*' for each sentence *A* of L^+. We follow Gupta and Belnap [5] in defining S $=_{df}$ {*A*: *A* is a sentences of L^+}. A classical model M = ⟨D, I⟩ for *L* is a *ground* model for *L* iff both S $\subseteq$ D, and I('*A*') = *A* for each *A* $\in$ S. A classical ground model for *L* is a representation of the supposedly unproblematic fragment of L^+: a representation of which terms refer to which objects, and of which objects have which nonsemantic properties.

We might want to extend a ground model M = ⟨D, I⟩ for L to a classical model M′ = ⟨D, I′⟩ for L^+ so that the extension of $\boldsymbol{T}$ in M′ (= {d ∈ D: I′($\boldsymbol{T}$)(d) = **t**}) is the set of sentences of L^+ true in M′: we will call such a model M′ a *Tarski* model. A Tarski model for L^+ is one in which truth behaves as it intuitively should. Unfortunately, some ground models cannot be extended to Tarski models. Suppose that the language L has one nonquote name, b, no function symbols, and no predicate or relation symbols. Also suppose that M = ⟨D, I⟩ is a ground model, where D = S and where I(b) = $\neg\boldsymbol{T}b$. Finally suppose that M′ = ⟨D, I′⟩ is a classical model extending M. Note that I′($\boldsymbol{T}$)($\neg\boldsymbol{T}b$) = **t** iff I′($\boldsymbol{T}$)(I′(b)) = **t** iff $\boldsymbol{T}b$ is true in M′ iff $\neg\boldsymbol{T}b$ is not true in M′. So the extension in M′ of $\boldsymbol{T}$ cannot be the set of sentences true in M′. This is simply a formalization of the liar's paradox. The paradox can be attributed both to the intuitively desired behaviour of truth—i.e. by our desire to extend M to a Tarski model—and the to fact, unproblematic in itself, that the name b refers to $\neg\boldsymbol{T}b$.

We can consider other problematic ground models. If I(b) = $\boldsymbol{T}b$ rather than $\neg\boldsymbol{T}b$, then we can extend M to two Tarski models, M′ = ⟨D, I′⟩ and M″ = ⟨D, I″⟩, where I′($\boldsymbol{T}$)($\boldsymbol{T}b$) = **t** and I″($\boldsymbol{T}$)($\boldsymbol{T}b$) = **f**. The problem with the truth-teller is that there seems to be nothing to decide between M′ and M″. Established terminology has it that the truth-teller is pathological, but not so bad as the liar as to be paradoxical.[1] The source of pathologicality need not be self-reference or even circular reference: Yablo [11] gives an example of a pathological ground model with no referential circularity. But even in Yablo's example, the pathology might be attributed to some kind of vicious reference in the ground model. Finally, the vicious reference in a ground model need not be *singular terms* viciously referring to sentences. Consider a language L with one unary predicate G, and a ground model M = ⟨D, I⟩ where D = S and I(G)(A) = **t** iff $A = \forall x(Gx \rightarrow \neg\boldsymbol{T}x)$ for every sentence A of L^+, so that the extension of G is $\{\forall x(Gx \rightarrow \neg\boldsymbol{T}x)\}$. M cannot be extended to a Tarski model. Here it is a predicate, G, that is viciously referring.

[1] Gupta pointed this out, in correspondence.

It has long been recognized that not all self-reference is vicious or pathology-producing. Consider,

The italicized sentence on page three contains four words.

This is an unpathologically false self-referential sentence. Formally, suppose that the language L has exactly one nonquote name, c, and one nonlogical predicate, G, and no function symbols. Also suppose that $M = \langle D, I \rangle$ is a ground model such that $I(c) = Gc$, and $I(G)(Gc) = \mathbf{f}$. Then, from an intuitive point of view, there is nothing vicious about the self-reference: Gc is unpathologically false, and reference to it should be non-vicious.

We have been toying with an intuitive notion of a *pathological* ground model and an intuitive notion of a ground model with *vicious reference*. These notions suggest the complementary notions of a *nonpathological* ground model, and of a ground model with *no vicious reference*. Gupta and Belnap [5] claim that one advantage of their particular approach to truth is "its consequence that truth behaves like an ordinary classical concept under certain conditions—conditions that can roughly be characterized as those in which there is no vicious reference in the language [i.e., in the ground model]." (p. 201) Aiming to clarify this remark, they define *Thomason* ground models, the ground models in which—from Gupta and Belnap's particular theoretic perspective—truth behaves like a classical concept.[2] They investigate some of the conditions under which a ground model is Thomason: by considering a number of examples and theorems, they build a case by case argument that a model is Thomason when there is no vicious reference. Though their notion of a Thomason model is formal and precise, their notion of "no vicious reference" remains informal and intuitive throughout, precluding a mathematical proof of their conclusion and necessitating the case by case argument.[3]

[2] Gupta and Belnap do not explicitly assert that the Thomason models are precisely those in which truth behaves classically, but it is clear from their discussion that they are intended as such.

[3] Gupta pointed this out, in correspondence.

In the current paper, we will approach the topic from a perspective slightly different to Gupta and Belnap's, but prompted by their discussion. We will consider a number of theories of truth, both Gupta and Belnap's *revision* theories, and theories motivated by the *fixed point semantics* in opposition to which Gupta and Belnap develop their approach. We will give a formal definition of when there is no vicious reference in a ground model *relative to this or that theory* **T**, and a formal definition of when truth behaves like a classical concept in a ground model M *relative to this or that theory* **T**.[4] We will then state a desideratum on any theory **T** of truth, whether a revision theory or a fixed point theory: If there is no vicious reference (relative to **T**) in a ground model M, then truth should behave like a classical concept (relative to **T**) in M. This echoes the "adequacy condition" in Gupta [4]—the paper which first introduces Thomason models—on any theory of truth: "For models M belonging to a certain class—a class that we have not formally defined but which in intuitive terms contains models that permit only benign kinds of self-reference—the theory should entail that all Tarski biconditionals are assertable in the model M." (p. 194) We will prove that some of the fixed point theories satisfy our desideratum, and that at least one the Gupta-Belnap revision theories does not.

If our desideratum were identical to the Gupta-Belnap desideratum—that truth behave like a classical concept in the absence of vicious reference—then the significance of our results would be clear: we would see that a number of rivals of the Gupta-Belnap theories share the advantage that Gupta and Belnap claim for their approach, and that at least one of Gupta and Belnap's revision theories does not. This would present a challenge to their suggestion that the satisfaction of the desideratum is an advantage that is distinctive of their approach, or at least a reason to qualify this suggestion. But, as Gupta has cautioned us in correspondence, not only is the Gupta-Belnap notion of non-vicious reference informal and intuitive, it is also theory-neutral while ours

[4] We will also consider interpreting the fixed point semantics so that the behaviour of truth is not determined by the ground model. We agree with M. Kremer [6] that Kripke [8] favours this interpretation.

is theory-relative. We will wait until our formal definitions are on the table before discussing these issues.

Our discussion will not be self-contained. In particular, we will rely on the notation and concepts introduced in §2 and §3 of the companion paper, Kremer [7]. We will also rely on other definitions from [7]: we will specifically and explicitly reference anything from [7] which is not from §2 or §3 of that paper.

§2. Fixed point semantics and revision theories of truth. In §2 and §3 of Kremer [7], we develop the fixed point semantics (Kripke [8] and Martin and Woodruff [9]) and the revision theoretic semantics (Gupta and Belnap [5]) for languages expressing their own truth concepts.

In the fixed-point semantics, a ground model M is expanded to a three-valued model M + h, where the *hypothesis* h is a three-valued interpretation of ***T***. The details depend on the *scheme of abbreviation* used to assign truth values to composite sentences. In [7] we introduced the weak Kleene scheme, μ; the strong Kleene scheme, κ; the supervaluation scheme, σ; and two variants on the supervaluation scheme, $\sigma 1$ and $\sigma 2$. Given a ground model M, each scheme ρ is associated with a certain *jump* operator ρ_M on hypotheses, h. The models M + h in which ***T*** plausibly expresses *truth* are those for which h is a *fixed point* of ρ_M. Two fixed points of ρ_M are of particular interest: the least fixed point, $\mathrm{lfp}(\rho_M)$, and the greatest intrinsic fixed point, $\mathrm{gifp}(\rho_M)$. (See [7], §2, for the details.) Thus, in [7], we define ten fixed-point theories of truth: the least fixed point theories $\mathbf{T}^{\mathrm{lfp},\, \mu}$, $\mathbf{T}^{\mathrm{lfp},\, \kappa}$, $\mathbf{T}^{\mathrm{lfp},\, \sigma}$, $\mathbf{T}^{\mathrm{lfp},\, \sigma 1}$, and $\mathbf{T}^{\mathrm{lfp},\, \sigma 2}$; and the greatest intrinsic fixed point theories $\mathbf{T}^{\mathrm{gifp},\, \mu}$, $\mathbf{T}^{\mathrm{gifp},\, \kappa}$, $\mathbf{T}^{\mathrm{gifp},\, \sigma}$, $\mathbf{T}^{\mathrm{gifp},\, \sigma 1}$; and $\mathbf{T}^{\mathrm{gifp},\, \sigma 2}$. According to each of these theories, a sentence *A* is *valid* in a ground model M iff *A* is true in the relevant least or greatest intrinsic fixed point.

In the revision-theoretic semantics, we do not expand a ground model M to any *particular* model M + h for the whole language. Rather, we consider sequences of classical hypotheses, h, generated by progressively revising some initial hypothesis. For example, if a liar sentence is originally hypothesized to be true, it will be hypothesized to be false, upon revision, and so on.

Some sentence will eventually stabilize. In [7], we define three revision theories $\mathbf{T}^{*}$, $\mathbf{T}^{\#}$, and $\mathbf{T}^{c}$. According to $\mathbf{T}^{*}$, a sentence *A* is *valid* in a ground model M iff *A* eventually stablizes as true, no matter hypothesis h we begin with. For $\mathbf{T}^{\#}$, we weaken the notion of stability, and for $\mathbf{T}^{c}$ we restrict our attention to particular well-behaved *revision sequences*. (See [7], §4, for details.)

Given a ground model M, each of these thirteen theories is determined by the verdict it gives, for each sentence *A* of L^{+}, on whether *A* is *valid* in M. As pointed out in [7], these theories all rely on what M. Kremer [6] calls the *supervenience of semantics*: the intuition that the interpretation of ***T*** should be determined by the interpretation of the nonsemantic names, function symbols and predicates, as represented by a ground model. M. Kremer argues both that that Kripke [8] does not endorse this proposal, and that this proposal misinterprets the fixed point semantics. In particular, many ground models M allow many fixed points, in each of which ***T*** has a distinct interpretation. If ***T*** only means truth if the language is spoken in accordance with a fixed point, and if the interpretation of ***T*** is to be determined by the ground model as on the supervenience proposal, then a given theory of truth should privilege some particular fixed point, e.g. the least fixed point. M. Kremer argues that the fixed point semantics is meant to formalize the fixed point conception of truth, according to which, as Kripke [8] puts it, "we are entitled to assert (or deny) of a sentence that it is true precisely under the circumstances when we can assert (or deny) the sentence itself." But this conception favours no particular fixed point. (See [7], §2, for more discussion of this point.)

In [7], we formalize the supervenience proposal, in the fixed point setting, by defining the theories $\mathbf{T}^{\mathrm{lfp},\,\rho}$ and $\mathbf{T}^{\mathrm{gifp},\,\rho}$ for each scheme of evaluation ρ. Each of these theories determine when a sentence is valid in a ground model: either when it is true in the least fixed point of ρ_{M} or in the greatest fixed point of ρ_{M}. When it comes to the nonsupervenience proposal, the primary notion should not be when a sentence is valid in a *ground* model, since a ground model does not fix the interpretation of the whole language. The most obvious analogous notion we have is as follows.

Definition 2.1. Suppose that M is a ground model and that h is a fixed point of μ_M, κ_M, σ_M, or any monotone operator on hypotheses. The sentence *A* of L^+ *is valid in the extended model* M + h iff h(*A*) = **t**.

This notion is not relativized to a theory. But whether the notion of validity in M + h is well-defined for a given h is relative to an evaluation scheme, since whether h is a fixed point is relative to an evaluation scheme. So we have a different nonsupervenience fixed point approach for each evaluation scheme.

§3. Truth behaving like a classical concept. We have not yet introduced the notions of "no vicious reference", nor of truth behaving like a classical concept. But consider a classical ground model M = ⟨D, I⟩ that makes no distinctions, other than with quote names, among the sentences of L^+: for an extreme case, suppose that *L* has no nonquote names, no function symbols and no nonlogical predicates. There seems to be no opportunity for vicious reference of any kind under these circumstances. And yet lfp(μ_M) and lfp(κ_M) are nonclassical: thus it seems that neither of the least fixed point theories $\mathbf{T}^{\mathrm{lfp},\,\mu}$ or $\mathbf{T}^{\mathrm{lfp},\,\kappa}$ dictates that truth behaves like a classical concept in M. This is a simple example of the kind of result that Gupta and Belnap find counterintuitive: despite the absence of vicious reference, truth does not behave like a classical concept on these least fixed point theories. (It is worth noting that lfp(σ_M) is classical in M, as follows from [7], Corollary 4.24.)

Gupta and Belnap introduce their notion of a *Thomason* model (see [7], Definition 4.7) in order to clarify the advantage that they claim for their approach: "its consequence that truth behaves like an ordinary classical concept under ... conditions that can roughly be characterized as those in which there is no vicious reference in the language."

Definition 3.2. ([5]) A ground model M is *Thomason* iff all τ_M-sequences culminate in one and the same fixed point.

The notion of a Thomason model is a formal articulation of the notion of a nonpathological ground model, i.e. a ground model in which truth behaves like a classical concept. Though

Thomason models are not defined in a theory-relative manner, they formalize nonpathologicality *from the revision-theoretic perspective*: from a least fixed point perspective, for example, nonpathologicality would be characterized in terms of the properties of the least fixed point of μ_M, κ_M, σ_M, $\sigma1_M$, $\sigma2_M$, or whatever, rather than in terms of the various possible τ_M-sequences.

Soon after introducing Thomason models, Gupta and Belnap note that for every Thomason model M, the classical jump operator τ_M has a unique fixed point, say h.[5] They point out that, if M is Thomason, then $A \in \mathbf{V}^*_M$ iff $A \in \mathbf{V}^\#_M$ iff h(A) = **t** for every sentence A of L^+. Upon this they remark, "So, both theories $\mathbf{T}^*$ and $\mathbf{T}^\#$ dictate that truth behaves like a classical concept in Thomason models." This suggests the following definition, also given in [7].

Definition 3.3. $\mathbf{T}^*$ [$\mathbf{T}^\#$, $\mathbf{T}^c$] *dictates that truth behaves like a classical concept in the ground model* M iff $A \in \mathbf{V}^*_M$ [$\mathbf{V}^\#_M$, $\mathbf{V}^c_M$] or $\neg A \in \mathbf{V}^*_M$ [$\mathbf{V}^\#_M$, $\mathbf{V}^c_M$] for every sentence A of L^+.

The analogous definition, also given in [7], for the fixed point theories is as follows.

Definition 3.4. Let $\rho = \mu$, κ, σ, or $\sigma1$ or $\sigma2$. $\mathbf{T}^{\text{lfp},\,\rho}$ [$\mathbf{T}^{\text{gifp},\,\rho}$] *dictates that truth behaves like a classical concept in the ground model* M iff $A \in \mathbf{V}^{\text{lfp},\,\rho}_M$ [$\mathbf{V}^{\text{gifp},\,\rho}_M$] or $\neg A \in \mathbf{V}^{\text{lfp},\,\rho}_M$ [$\mathbf{V}^{\text{gifp},\,\rho}_M$] for every sentence A of L^+.

On any nonsupervenience fixed point approach, the issue is not whether truth behaves like a classical concept in a *ground* model, but rather in an *extended* model.

Definition 3.5. *Truth behaves like a classical concept in the extended model* M + h iff h is a classical fixed point of τ_M.

Theorem 3.6. A ground model is Thomason iff $\mathbf{T}^*$ dictates that truth behaves like a classical concept in it.

Proof. This follows immediately from the definitions. ⊣

§4. No vicious reference. Gupta [4] tentatively suggests that non-vicious reference can be defined via the notion of a Thomason model. But the immediacy of the proof of Theorem 3.6 suggests that the notion of a Thomason ground model is a notational variant of the notion of a

[5] τ_M having a unique fixed point is a necessary but not a sufficient condition for M to be Thomason.

ground model in which $\mathbf{T}^*$ dictates that truth behaves like a classical concept. For non-vicious reference, we would like a characterization that is linked to the referential behaviour of the names and predicates in the ground model. Gupta [4] suggests that whether reference is non-vicious in a ground model M is related to what distinctions can be made in M among the sentences of L^+: "Now, what sorts of self-reference can we allow in *L*? What kinds of distinctions among the sentences containing the truth predicate can we make without violating the fundamental intuition?"[6] (p. 191) This second question is made more precise with Gupta and Belnap's notion of a name's, predicate's, or function symbol's interpretation being *neutral* relative to some subset X of the domain of discourse D. See Definition 4.4, immediately below. (Gupta and Belnap's [5] notion of X-neutrality generalizes Gupta's [4] notion of S-neutrality.)

Definition 4.3. ([3], Definitions 2D.2) Suppose that $M = \langle D, I\rangle$ is a model for *L* and $X \subseteq D$.

(i) The interpretation of a name *c* is X-*neutral* in M iff $I(c) \notin X$.

(ii) The interpretation of an n-place predicate *F* is X-*neutral* in M, iff for all $d_1, ..., d_n, d'_i \in D$, if $d_i, d'_i \in X$ then $I(F)(d_1, ..., d_i, ..., d_n) = I(F)(d_1, ..., d'_i, ..., d_n)$.

(iii) The interpretation of an n-place function symbol *f* is X-*neutral* in M, iff both the range of I(*f*) is disjoint from X and for all $d_1, ..., d_n, d'_i \in D$, if $d_i, d'_i \in X$ then $I(f)(d_1, ..., d_i, ..., d_n) = I(f)(d_1, ..., d'_i, ..., d_n)$.

Definition 4.4. ([3], Definition 6A.2) A model $M = \langle D, I\rangle$ is X-*neutral* iff the interpretations in M of all the nonquote names, nonlogical predicates, and function symbols are X-neutral.

Gupta and Belnap prove a number of theorems relating a ground model's ability to make distinctions among the sentences of L^+ to its Thomasonness—i.e. to whether $\mathbf{T}^*$ dictates that truth behaves like a classical concept in it. They begin as follows. (See [5] or [7] for proofs.)

[6] The "fundamental intuition" about truth is that "from any sentence *A* the inference to another sentence that asserts that *A* is true is warranted. And conversely: from the latter sentence the inference to *A* is also warranted." (Gupta [4], p. 181).

Theorem 4.5. ([5], Theorem 6A.5) If the ground model M is S-neutral then all τ_M-sequences culminate in the same fixed point, i.e., M is Thomason.

Theorem 4.6. ([5], Theorem 6B.4, Convergence to a fixed point I) If M is X-neutral then M is Thomason, provided that X contains either (i) all sentences that have occurrences of $\boldsymbol{T}$, or (ii) all sentences that are μ-ungrounded in M, or (iii) all sentences that are κ-ungrounded in M, or (iv) all sentences that are σ-ungrounded in M.

Gupta and Belnap present the following example as an application of Theorem 4.6.

Example 4.7. ([5], Example 6B.6) Suppose that the ground model M = ⟨D, I⟩ is S-neutral except for the name *a*. Furthermore suppose that *Hb* is true in M. Then M is Thomason if (i) I(*a*) = *Hb*, (ii) I(*a*) = ***T***‘*Hb*’, (iii) I(*a*) = *Hb* ∨ ¬***T****a*, or (iv) I(*a*) = ***T****a* ∨ ¬***T****a*.

Theorems 4.5 and 4.6 and Example 4.7 are part of a case by case argument that if there is no vicious reference in the ground model M—if no vicious distinctions can be made in M among the sentences of L^+—then M is Thomason. Given Theorem 3.6, this can be interpreted as a case by case argument that $\mathbf{T}^*$ satisfies the following *Gupta-Belnap Desideratum* on theories **T** of truth (see Gupta’s [4] "adequacy condition" cited in §1 above):

Gupta-Belnap Desideratum. *If there is no vicious reference in the ground model* M *then* **T** *dictates that truth behaves like a classical concept in* M.

As mentioned in §1, both this desideratum and the argument that the Gupta-Belnap theory $\mathbf{T}^*$ satisfies it rely on an informal, intuitive, theory-neutral notion of non-vicious reference.[7]

The general strategy implicit in Theorems 4.5 and 4.6 and Example 4.7 is as follows. Find some set Y of intuitively unproblematic sentences. If M is (S - Y)-neutral—if M cannot make any distinctions among potentially problematic sentences—then M is Thomason. In Theorem 4.5, Y = ∅. In Theorem 4.6, Y can be any set of ***T***-free sentences, μ-grounded sentences, κ-grounded sentences, or σ-grounded sentences. In Example 4.7, Y = {*Hb*}, {***T***‘*Hb*’}, {*Hb* ∨ ¬***T****a*}, or {***T****a* ∨ ¬***T****a*}. There are fairly strong theory-neutral intuitions that, in each of these cases, reference

[7] This desideratum is not quite explicit in either [5] or [4], but we take it to be implicit.

to the objects in Y is non-vicious. And if one of these is a case of vicious reference, it would present no counterexample to the claim that $\mathbf{T}^{*}$ satisfies the Gupta-Belnap desideratum. (It would present a counterexample to the claim that $\mathbf{T}^{*}$ satisfies a converse desideratum.)

Interest in theories other than $\mathbf{T}^{*}$ might prompt a similar case by case consideration of whether our other theories dictate that truth behaves like a classical concept when there is no vicious reference. Regarding $\mathbf{T}^{\mathrm{lfp},\, \mu}$ and $\mathbf{T}^{\mathrm{lfp},\, \kappa}$, we have a negative result: these two theories never dictate that truth behaves like a classical concept (See the proof of [7], Theorem 4.5). For a large number of other theories, we get the following analogues of Theorems 4.5 and 4.6. (These follow from [7], Theorem 4.21 and Corollary 4.24.)

Theorem 4.8. If the ground model M is S-neutral then $\mathbf{T}^{*}$, $\mathbf{T}^{\#}$, $\mathbf{T}^{c}$, $\mathbf{T}^{\mathrm{lfp},\, \sigma}$, $\mathbf{T}^{\mathrm{lfp},\, \sigma 1}$, $\mathbf{T}^{\mathrm{lfp},\, \sigma 2}$, and $\mathbf{T}^{\mathrm{gifp},\, \rho}$ ($\rho = \mu, \kappa, \sigma, \sigma 1$, or $\sigma 2$) dictate that truth behaves like a classical concept in M.

Theorem 4.9. Suppose that the ground model M is X-neutral where X contains either (i) all sentences that have occurrences of ***T***, or (ii) all sentences that are μ-ungrounded in M, or (iii) all sentences that are κ-ungrounded in M, or (iv) all sentences that are σ-ungrounded in M. Then $\mathbf{T}^{*}$, $\mathbf{T}^{\#}$, $\mathbf{T}^{c}$, $\mathbf{T}^{\mathrm{lfp},\, \sigma 1}$, $\mathbf{T}^{\mathrm{lfp},\, \sigma 2}$, and $\mathbf{T}^{\mathrm{gifp},\, \rho}$ ($\rho = \mu, \kappa, \sigma, \sigma 1$, or $\sigma 2$) dictate that truth behaves like a classical concept in M. In all cases but (iv), $\mathbf{T}^{\mathrm{lfp},\, \sigma}$ dictates that truth behaves like a classical concept in M.

We have a negative result for $\mathbf{T}^{\mathrm{lfp},\, \sigma}$. (See [7], Example 5.12.)

Theorem 4.10. For some X-neutral ground model M, where X contains all sentences that are σ-ungrounded in M, $\mathbf{T}^{\mathrm{lfp},\, \sigma}$ does not dictate that truth behaves like a classical concept in M.

So far the informal case by case argument has ruled out $\mathbf{T}^{\mathrm{lfp},\, \mu}$ and $\mathbf{T}^{\mathrm{lfp},\, \kappa}$ as satisfying the Gupta-Belnap desideratum, and has ruled out $\mathbf{T}^{\mathrm{lfp},\, \sigma}$ on the assumption that the conditions on M in Theorem 4.10 represent non-vicious reference. But we still have an informal argument that a large number of our theories satisfy the desideratum, in particular $\mathbf{T}^{*}$, $\mathbf{T}^{\#}$, $\mathbf{T}^{c}$, $\mathbf{T}^{\mathrm{lfp},\, \sigma 2}$ and $\mathbf{T}^{\mathrm{gifp},\, \rho}$ where $\rho = \mu, \kappa, \sigma, \sigma 1$, or $\sigma 2$. The following is Gupta and Belnap's last theorem relating (S - Y)-

neutrality to Thomasonness—i.e. their last theorem in their case by case argument that $\mathbf{T}^*$ satisfies the Gupta-Belnap desideratum. For a proof, see [5] or [7].

Theorem 4.11. ([5], Theorem 6B.8, Convergence to a fixed point II) Suppose that M is an (S - Y)-neutral model and that Y contains only sentences that are either stably **t** in all τ_M-sequences or stably **f** in all such sequences—in other words, $Y \subseteq \{A\colon\ A \in \mathbf{V}^*_M \text{ or } \neg A \in \mathbf{V}^*_M\}$. Then, M is Thomason.

Given Theorem 4.9, above, we can immediately strengthen Theorem 4.11.

Theorem 4.12. Suppose that M is an (S - Y)-neutral model where $Y \subseteq \{A\colon\ A \in \mathbf{V}^*_M \text{ or } \neg A \in \mathbf{V}^*_M\}$. Then the following theories dictate that truth behaves like a classical concept in M: $\mathbf{T}^*$, $\mathbf{T}^\#$, $\mathbf{T}^c$, and $\mathbf{T}^{\text{gifp}, \rho}$, for $\rho = \mu, \kappa, \sigma, \sigma 1$, or $\sigma 2$.

On the negative side, we have the following. For a proof, see [5], Example 6B.13.

Theorem 4.13. There are (S - Y)-neutral models where $Y \subseteq \{A\colon\ A \in \mathbf{V}^*_M \text{ or } \neg A \in \mathbf{V}^*_M\}$ and in which neither $\mathbf{T}^{\text{lfp}, \sigma}$, $\mathbf{T}^{\text{lfp}, \sigma 1}$, nor $\mathbf{T}^{\text{lfp}, \sigma 2}$ dictates that truth behaves like a classical concept.

The conditions on the ground model in Theorems 4.5 and 4.6 might be clear cases of non-vicious reference, from an intuitive theory-neutral perspective. But the condition on the ground model M in Theorems 4.11 and 4.13 is that M be (S - Y)-neutral model, where $Y \subseteq \{A\colon A \in \mathbf{V}^*_M$ or $\neg A \in \mathbf{V}^*_M\}$. Thus, on the strategy we suggested for interpreting Theorems 4.5 and 4.6, the set Y of unproblematic sentences can be any subset of $\{A\colon\ A \in \mathbf{V}^*_M \text{ or } \neg A \in \mathbf{V}^*_M\}$. At this point in their case by case argument, Gupta and Belnap no longer seem to be working with clearly theory-neutral intuitions concerning non-vicious reference: the intuitions at work are intuitions that rate as non-vicious reference to the sentences in $\{A\colon\ A \in \mathbf{V}^*_M \text{ or } \neg A \in \mathbf{V}^*_M\}$. This seems motivated by the revision theoretic semantics, and more specifically by the theory $\mathbf{T}^*$.

It does not affect Gupta and Belnap's argument for the conclusion that truth always behaves classically in the absence of vicious reference, if truth occasionally behaves classically in the presence of vicious reference. But we might want to interpret Theorem 4.13 as evidence that $\mathbf{T}^{\text{lfp}, \sigma}$, $\mathbf{T}^{\text{lfp}, \sigma 1}$, and $\mathbf{T}^{\text{lfp}, \sigma 2}$ do not satisfy the Gupta-Belnap desideratum, in which case we want to be pretty sure that the condition placed on M in the statement of the theorem is a condition under which there is no vicious reference. We contend that from a perspective not already informed by revision theory, we simply cannot be sure of this.

We grant that there are informal theory-neutral intuitions about vicious reference: the reference involved in the liar is certainly vicious, and the reference involved in the truth-teller is almost certainly vicious. For an illustrative intuitive case of non-vicious reference, suppose that M = $\langle$D, I$\rangle$ is a ground model and that *c* is a name and *G* is a classical unary predicate. Further, suppose that I(*c*) = ***T***‘*Gc*’. The sentence ***T***‘*Gc*’ does not have any truth-value in the ground model, since I assigns no extension or anti-extension to ***T***. But on any *reasonable* theory of truth, ***T***‘*Gc*’ will behave classically, and will be assigned the same classical truth-value as *Gc*. So reference to ***T***‘*Gc*’ ought to be non-vicious.

Here, we already have an implicit relativization of non-vicious reference to the theory of truth: reference to ***T***‘*Gc*’ should be non-vicious on any *reasonable* theory of truth since ***T***‘*Gc*’ should behave classically on any *reasonable* theory of truth. Consider an unreasonable theory of truth, $\mathbf{T}^{\mathrm{null}}$, which assigns to every sentence of the form ***Tb*** the truth-value **n**. According to $\mathbf{T}^{\mathrm{null}}$, it is not so clear whether reference to ***T***‘*Gc*’ is vicious: after all, according to $\mathbf{T}^{\mathrm{null}}$, this sentence does not behave classically. Our informal intuitions about what kind of reference is non-vicious are informed by our informal intuitions about what sentences will behave classically in *reasonable* theories of truth. Thus even our informal intuitions are, in some sense, theory-relative: relative to reasonable theories.

We might want a tool for a more fine-grained comparison of theories. The most general *formal* articulation of non-vicious reference, we suggest, will be theory-relative: non-vicious reference will be reference to non-sentences or to unproblematic sentences, i.e. sentences that get a definite, stable, and classical truth-value, a theory-relative matter. Extending this to predicates, non-vicious distinctions will be distinctions among the non-sentences together with the unproblematic sentences.

Definition 4.15. Let $\rho = \mu, \kappa, \sigma, \sigma 1$ or $\sigma 2$. Let $\mathbf{T} = \mathbf{T}^{*}, \mathbf{T}^{\#}, \mathbf{T}^{c}, \mathbf{T}^{\mathrm{lfp},\, \rho}$, or $\mathbf{T}^{\mathrm{gifp},\, \rho}$, with $\mathbf{V} = \mathbf{V}^{*}_{\mathrm{M}}$, $\mathbf{V}^{\#}_{\mathrm{M}}$, $\mathbf{V}^{c}_{\mathrm{M}}$, $\mathbf{V}^{\mathrm{lfp},\, \rho}_{\mathrm{M}}$, or $\mathbf{V}^{\mathrm{gifp},\, \rho}_{\mathrm{M}}$, corresponding to $\mathbf{T}$. Let M be a ground model for *L*. **T** *dictates that there is no vicious reference in* M iff M is (S - Y)-neutral for some $\mathrm{Y} \subseteq \{A\colon A \in \mathbf{V}$ or $\neg A \in \mathbf{V}\}$.

The analogue on the nonsupervenience fixed point approaches is as follows.

Definition 4.16. Suppose that M is a ground model and that h is a fixed point of μ_{M}, κ_{M}, σ_{M}, $\sigma 1_{\mathrm{M}}$, $\sigma 2_{\mathrm{M}}$, or any monotone operator on hypotheses. *There is no vicious reference in the extended model* M + h iff M is (S - Y)-neutral where $\mathrm{Y} \subseteq \{A\colon\ \mathrm{h}(A) = \mathbf{t}$ or $\mathrm{h}(A) = \mathbf{f}\}$.

§5. No vicious reference, again. In §4, we took up Gupta's [4] suggestion that whether reference is vicious is related to what kind of distinctions can be made within the domain of discourse. In case this strategy is unintuitive, we will advance another way to develop a notion of non-vicious reference. Beginning with names, we will take all quote names to refer non-viciously, and we will take a nonquote name to refer non-viciously if it refers to either a non-sentence or an unproblematic sentence. Of course, this is a theory-relative issue.

For predicates, we begin with classical one-place predicates. Each name is closely related to a unary predicate, as "Pegasus" is to "pegasizes". Suppose that $\mathrm{M} = \langle \mathrm{D}, \mathrm{I} \rangle$ is a ground model, where the extension of *G* is $\{\mathrm{I}(b)\}$ and the extension of *H* is $\{\mathrm{I}(c)\}$. Intuitively, if the names *b* and *c* refer non-viciously then so do the predicates *G* and *H*. We should not place any special emphasis on the extension as opposed to the anti-extension of a predicate: a classical predicate's signification is just as much determined by its anti-extension as its extension. One way to think of a classical predicate's signification is as the way it partitions the universe of discourse. So it seems intuitive to say that $\neg G$ and $\neg H$ also refer non-viciously. It also seems intuitive to say that $(G \vee H)$, with the extension $\{\mathrm{I}(b), \mathrm{I}(c)\}$, refers non-viciously.

Generalizing, for classical unary predicates we get the following: the predicate *P refers non-viciously* iff either every object in *P*'s extension is unproblematic or every object in *P*'s anti-extension is unproblematic, where the unproblematic objects are the nonsentences and unproblematic sentences. Note that referring unproblematically is closed under Boolean operators. To extend this to n-place classical predicates, we will say that an ordered n-tuple $\langle d_1, ..., d_n\rangle \in D^n$ is *unproblematic* iff each of $d_1, ..., d_n$ is. Then we can say that the n-place predicate *R refers non-viciously* iff either every n-tuple in *R*'s extension is unproblematic, or every n-tuple in *R*'s anti-extension is unproblematic.

For n-ary function symbols, we combine the strategy for nonquote names and for predicates, in Definition 5.1, below. We will assume that *L* is a language with quote names for the sentences in L^+, and that $\mathbf{T} = \mathbf{T}^*, \mathbf{T}^\#, \mathbf{T}^c, \mathbf{T}^{lfp,\, \rho}$, or $\mathbf{T}^{gifp,\, \rho}$ with $\mathbf{V} = \mathbf{V}^*_M, \mathbf{V}^\#_M, \mathbf{V}^c_M, \mathbf{V}^{lfp,\, \rho}_M$, or $\mathbf{V}^{gifp,\, \rho}_M$, corresponding to **T**.

Definition 5.1. Suppose that M = $\langle D, I\rangle$ is a classical model for *L*.

(i) **T** *dictates that the nonquote name b refers non-viciously in* M iff $I(b) \in (D - S) \cup \{A\text{: } A \in \mathbf{V} \text{ or } \neg A \in \mathbf{V}\}$.

(ii) **T** *dictates that the nonlogical n-ary predicate R refers non-viciously in* M iff either (a) for every $\langle d_1, ..., d_n\rangle$ in the extension of R, each $d_i \in (D - S) \cup \{A\text{: } A \in \mathbf{V} \text{ or } \neg A \in \mathbf{V}\}$; or (b) for every $\langle d_1, ..., d_n\rangle$ in the antiextension of R, each $d_i \in (D - S) \cup \{A\text{: } A \in \mathbf{V} \text{ or } \neg A \in \mathbf{V}\}$.

(iii) **T** *dictates that the n-ary function symbol f refers non-viciously in* M iff both (1) for every $d_1, ..., d_n \in D$, $I(f)(d_1, ..., d_n) \in (D - S) \cup \{A\text{: } A \in \mathbf{V} \text{ or } \neg A \in \mathbf{V}\}$; and (2) for each $d \in D$, either (2a) for every n-tuple $\langle d_1, ..., d_n\rangle$ such that $I(f)(d_1, ..., d_n) = d$, each $d_i \in (D - S) \cup \{A\text{: } A \in \mathbf{V} \text{ or } \neg A \in \mathbf{V}\}$; or (2b) for every n-tuple $\langle d_1, ..., d_n\rangle$ such that $I(f)(d_1, ..., d_n) \neq d$, each $d_i \in (D - S) \cup \{A\text{: } A \in \mathbf{V} \text{ or } \neg A \in \mathbf{V}\}$.

Definition 5.2. **T** *dictates that there is no vicious references in the ground model* M iff **T** dictates that every nonquote name, every nonlogical predicate and every function symbol refers non-viciously in M.

The analogue on the nonsupervenience fixed point approaches is as follows.

Definition 5.3. Suppose that M is a ground model and that h is a fixed point of μ_M, κ_M, σ_M, $\sigma 1_M$, $\sigma 2_M$, or any monotone operator on hypotheses. M + h *dictates that the name b, the n-ary* *predicate R, or the n-ary function symbol f, refers non-viciously* is defined as in Definition 6.1, with $\mathbf{T}$ replaced by M + h and with $\mathbf{V} = \{A\text{: } \mathrm{h}(A) = \mathbf{t} \text{ or } \mathrm{h}(A) = \mathbf{f}\}$. *There is no vicious reference in the extended model* M + h iff M + h dictates that every nonquote name, every nonlogical predicate other than $\boldsymbol{T}$ and every function symbol refers non-viciously.

Theorem 5.4. Definition 5.2 is equivalent to Definition 4.15, and Definition 5.3 is equivalent to Definition 4.16 (assuming, in all cases, that the ground model is classical).

Proof. Directly from the definitions. ⊣

§6. Does truth behave like a classical concept when there is no vicious reference? We are now ready to state our formal desideratum on supervenience theories of truth.

Desideratum 6.1. *If* $\mathbf{T}$ *dictates that there is no vicious reference in the ground model* M, *then* $\mathbf{T}$ *dictates that truth behaves like a classical concept in the ground model* M.

From the nonsupervenience fixed point perspective, we get a desideratum not on theories of truth, but on schemes of evaluation $\rho = \mu, \kappa, \sigma, \sigma1$ and $\sigma2$, which correspond to distinct nonsupervenience fixed point approaches.

Desideratum 6.2. *For every ground model* M *and every fixed point h of* ρ_{M}, *if there is no vicious reference in the extended model* M + h *then truth behaves like a classical concept in the extended model* M + h.

Our main results are as follows.

Theorem 6.3. (i) $\mathbf{T}^{*}$, $\mathbf{T}^{c}$, $\mathbf{T}^{\mathrm{lfp},\,\sigma2}$, and $\mathbf{T}^{\mathrm{gifp},\,\rho}$ satisfy Desideratum 6.1, for $\rho = \mu, \kappa, \sigma, \sigma1$, or $\sigma2$. (ii) $\mathbf{T}^{\#}$, $\mathbf{T}^{\mathrm{lfp},\,\mu}$, $\mathbf{T}^{\mathrm{lfp},\,\kappa}$, $\mathbf{T}^{\mathrm{lfp},\,\sigma}$, and $\mathbf{T}^{\mathrm{lfp},\,\sigma1}$ do not.

Theorem 6.4. (i) $\sigma2$ satisfies Desideratum 6.2, but (ii) μ, κ, σ and $\sigma1$ do not.

Theorems 6.3 and 6.4 (ii) follow from [7], Theorem 4.21 (2) and (3). The proof of Theorem 6.4 (i) is similar to the proof in [7] of Theorem 4.21 (2) for $\mathbf{T}^{\mathrm{lfp},\,\sigma2}$.

We recall Gupta's caution (§1): Desideratum 6.1 must be distinguished from the Gupta-Belnap desideratum, for which the notion of non-vicious reference is theory-neutral rather than theory-relative, and informal and intuitive rather than formal. Theorem 6.3 brings out a striking difference between Desideratum 6.1 and the Gupta-Belnap desideratum. Say that $\mathbf{T}' \geq_1 \mathbf{T}$ iff for every language L every ground model M and every sentence A of L^+, if A is valid in M according to $\mathbf{T}$ then A is valid in M according to $\mathbf{T}'$. (This is [7], Definition 4.1.) As Gupta has noted in correspondence, if a theory $\mathbf{T}$ satisfies the Gupta-Belnap desideratum then any theory $\mathbf{T}' \geq_1 \mathbf{T}$ is also bound to satisfy it. Not so with Desideratum 6.1, which is satisfied by $\mathbf{T}^*$ but not by $\mathbf{T}^\#$, although $\mathbf{T}^\# \geq_1 \mathbf{T}^*$. Consider the following example, which proves that $\mathbf{T}^\#$ does not satisfy Desideratum 6.1.

Example 6.5. ([5], Example 6B.9, and [7], Example 5.9) Consider a language L with no nonquote names, no function symbols, a one-place predicate G, and no other nonlogical predicates. Let L^+ be L extended with a one-place predicate $\boldsymbol{T}$, and suppose that L has a quote name 'C' for every sentence C of L^+. Let $A = \exists x(Gx \;\&\; \neg \boldsymbol{T}x)$, let $B = \exists x \exists y(Gx \;\&\; Gy \;\&\; \neg \boldsymbol{T}x \;\&\; \neg \boldsymbol{T}y \;\&\; x \neq y)$ and let $\mathrm{Y} = \{\boldsymbol{T}^n A\text{:}\;\; n \geq 0\}$. Let $\mathrm{M} = \langle \mathrm{D}, \mathrm{I}\rangle$ be the ground model where D is the set of sentences of L^+ and where $\mathrm{I}(G)(\mathrm{d}) = \mathbf{t}$ iff $\mathrm{d} \in \mathrm{Y}$. Note that every sentence in Y is *nearly* stably $\mathbf{t}$ in every τ_M-sequence, though not *stably* $\mathbf{t}$ in any τ_M-sequence. So $C \in \mathbf{V}^\#_\mathrm{M}$, for all $C \in \mathrm{Y}$. So $\mathbf{T}^\#$ dictates that there is no vicious reference in M. But $\mathbf{T}^\#$ does not dictate that truth behaves like a classical concept in M: as shown in [7], there is a τ_M-sequences $\boldsymbol{S}$ in which the sentence B is neither stably $\mathbf{t}$ nor stably $\mathbf{f}$. ⊣

If we accept Gupta and Belnap's informal case by case argument that $\mathbf{T}^*$ satisfies the Gupta-Belnap desideratum (see §5), then we must also accept that $\mathbf{T}^\#$ satisfies it. In that case the ground model in Example 6.5 must contain vicious reference. Does it? Since the language has no names, the question becomes whether we can use the predicate G to make vicious distinctions among the objects in the domain of discourse. Can we? There are *unstable* sentences that we can distinguish with G: $\mathrm{I}(G)(A) = \mathbf{t}$ and $\mathrm{I}(G)(B) = \mathbf{f}$, although each of A and B is unstable in some τ_M-sequence. But to take this to be a vicious distinction is to favour the notion of stability over the notion of near stability: we simply cannot use G to distinguish among sentences that fail to be nearly stable. From a perspective that favours neither stability nor near stability—i.e., neither $\mathbf{T}^*$ nor $\mathbf{T}^\#$—we maintain that the question whether the ground model in Example 6.5 has vicious reference is simply too imprecise to have a determinate answer.

But in this ground model, neither $\mathbf{T}^\#$ nor $\mathbf{T}^*$ dictates that truth behaves like a classical concept. So the question of whether $\mathbf{T}^\#$ or $\mathbf{T}^*$ satisfies the Gupta-Belnap desideratum is itself too imprecise to have a determinate answer. In particular, the Gupta-Belnap question of whether or not a model is Thomason when there is no vicious reference (understood informally) is too imprecise to have an answer: Example 6.5 is of a non-Thomason model which is a borderline case of non-vicious reference.

The best we can show for a theory **T**, using the informal notion of non-vicious reference, is that **T** satisfies what we will call the *weak* Gupta-Belnap desideratum: in *clear intuitive cases* of non-vicious reference, **T** dictates that truth behaves like a classical concept. But if we want to ask in general whether **T** dictates that truth behaves like a classical concept in the absence of vicious reference, and if we want our general question to be precise enough to have a definite answer, we are going to need a precise notion of non-vicious reference.

The advantage that Gupta and Belnap claim for their approach—the satisfaction of the Gupta-Belnap desideratum—is an imprecise advantage. When we make it precise, we discover that a number of the supervenience fixed point theories share the advantage (Theorem 6.3 (i)). And if we opt for a nonsupervenience fixed point approach, then the analogous advantage is had as long as we use the evaluation scheme σ2 (Theorem 6.4). Furthermore, Gupta and Belnap's revision theory $\mathbf{T}^\#$ does not share the precise version of the advantage (Theorem 6.3 (ii)). We believe that Gupta and Belnap should reconsider the place of their no-vicious-reference-implies-truth-behaves-classically intuition, since the precise version of this intuition is not satisfied by all of the revision theories and is satisfied by a number of the revision theories' rivals.

§7. Concluding remarks. Gupta and Belnap present the satisfaction of the no-vicious-reference-implies-truth-behaves-classically desideratum as one advantage of their approach to truth. We have formalized their desideratum, using a formal theory-relative notion of non-vicious reference rather than their informal theory-neutral notion. The revision theories $\mathbf{T}^{*}$ and $\mathbf{T}^{c}$ both have the advantage in its formalized form, but the revision theory $\mathbf{T}^{\#}$ does not. The supervenience fixed point theories $\mathbf{T}^{\mathrm{lfp},\, \sigma 2}$, $\mathbf{T}^{\mathrm{gifp},\, \mu}$, $\mathbf{T}^{\mathrm{gifp},\, \kappa}$, $\mathbf{T}^{\mathrm{gifp},\, \sigma}$, $\mathbf{T}^{\mathrm{gifp},\, \sigma 1}$, and $\mathbf{T}^{\mathrm{gifp},\, \sigma 2}$ have this advantage. And the nonsupervenience fixed point approach based on $\sigma 2$ has an analogous advantage. Although our desideratum is slightly different to Gupta and Belnap's, we believe that our results, at the very least, show that a number of fixed point theories are as attractive as the revision theories in regards to the behaviour of truth in the absence of vicious reference, and that at least one natural revision theory is suspect in this regard. We will consider a number of responses to our results.

Response 1. One response is to insist that we have warped the intuitive notion of non-vicious reference by formalizing it as we have, causing a resulting shift away from the original desideratum to something quite different. In reply, we could take the soft line that our desideratum on theories is merely an *alternative* desideratum to the Gupta-Belnap desideratum, and that the satisfaction of our desideratum should be seen as an alternative advantage to the satisfaction of theirs. We could also take a hard line, according to which their desideratum is so imprecise that the question of whether $\mathbf{T}^{\#}$ or even $\mathbf{T}^{*}$ satisfies it has no determinate answer (see §6, above). On this line, our desideratum can be seen as an appropriate precisification of theirs, and as a better device for comparing theories of truth. One way to combat the hard line would be to develop a formal, intuitively appealing, but theory-neutral notion of non-vicious reference, and to test whether our theories satisfy the new resulting desideratum.

Our tentative belief is that Gupta and Belnap's informal notion of non-vicious reference splits into a variety of notions upon formalization, one for each theory of truth. By way of analogy, consider an informal notion of logical consequence for a second order language. Under the pressure of formalization, this notion splits into a model-theoretic relation $\models$ between premise-sets and conclusions, and a proof-theoretic relation $\vdash$ (in fact, a number of proof-theoretic notions, depending on our choice of comprehension axiom). Certain desiderata involving the informal notion of logical consequence might be satisfied by $\vDash$ and others by $\vdash$. Similarly, different desiderata might be satisfied by ground models with no vicious reference according to **T** and according to **T′**.

The question might arise whether $\vDash$ or $\vdash$ is the *correct* formalization of second order logical consequence. Similarly, the question might arise whether one of our formal notions of non-vicious reference is the correct formalization of the intuitive notion. One answer might be that the correct formalization is the one generated by the correct theory of truth, maybe $\mathbf{T}^{\text{gifp}, \kappa}$ or $\mathbf{T}^{\#}$. Be that as it may, for each theory **T**, we still want to establish whether **T** satisfies the no-vicious-reference-implies-truth-behaves-classically intuition by using **T**'s own notion of no vicious reference.

Response 2. It is important to keep in mind that the satisfaction of such a desideratum does not constitute the most basic argument against fixed point theories and in favour of revision theories. Gupta and Belnap spend twenty-eight pages presenting quite different considerations against fixed point approaches to truth. Furthermore they motivate their revision theory quite independently of the no-vicious-reference-implies-truth-behaves-classically intuition. The satisfaction of this intuition is presented as an important bonus, but as a bonus of an otherwise motivated approach.

Response 3. It is worth noting that those fixed point theories that satisfy our desideratum are otherwise less appealing than those that do not. All of the greatest intrinsic fixed point theories satisfy our desideratum. But, although the gifp's has seemed a natural candidate for special attention, nowhere in the literature do we see arguments that any gifp delivers the correct interpretation of truth. The only lfp theory, among those we have considered, that satisfies our desideratum is $\mathbf{T}^{\text{lfp}, \sigma 2}$. And the only *nonsupervenience* fixed point approach that satisfies the analogous desideratum is the one that relies on the strongly consistent supervaluation scheme $\sigma 2$. $\mathbf{T}^{\text{lfp}, \sigma 2}$ and $\sigma 2$ have had little explicit attention in the literature, and few advocates. It is, however, worth noting that a number of constructions and remarks in McGee [10] depend on $\sigma 2$ and favour $\sigma 2$ over $\sigma 1$ and σ.[8]

Response 4. One might argue that truth only *genuinely* behaves like a classical concept in a ground model M when M is Thomason. When and only when M is Thomason, can we say, "when we revise a hypothetical extension ... for 'true' by repeated applications of τ_M, we find that ... we reach a stage after which the revision rule ceases to revise . Further, *no matter with what hypothesis we choose to initiate the revision process, we end up in the same fixed point*." (Gupta and Belnap [5], p. 134) This makes the Thomason models look privileged independently of their connection to the theory $\mathbf{T}^*$, i.e. independently of the fact that a ground model M is Thomason iff $\mathbf{T}^*$ dictates that truth behaves like a classical concept. An advocate of $\mathbf{T}^{\mathrm{lfp},\, \sigma 2}$ or some other fixed point theory might, however, remain unimpressed, arguing as follows: only if you are already committed to the view that the class of τ_M-sequences represents the behaviour of truth in M, will you want to privilege the class of ground models M such that all τ_M-sequences culminate in the same fixed point. If, on the other hand, the behaviour of truth in M is represented by $\mathrm{lfp}(\sigma 2_M)$, then whether or not all τ_M-sequences culminate in the same fixed point seems considerably less significant.

Response 5. Finally, we point out that a certain kind of fixed point theorist might be willing to jettison the no-vicious-reference-implies-truth-behaves-classically intuition altogether, at least when non-vicious reference is understood as we have been understanding it. On certain anaphoric analyses of truth,[9] a sentence of the form ***Tb*** inherits its semantic content, however such content is understood, from whatever sentence is referred to by *b*, whether *b* is a quote name or a nonquote name. *b* might refer to the sentence ***Tc***, so that ***Tb*** ultimately inherits its semantic content from whatever sentence is referred to by *c*. And so on.

[8] McGee refers to σ, $\sigma 1$, and $\sigma 2$ as σ_1, σ_2 and σ_3, respectively.

[9] See Grover, Camp and Belnap [3], Grover [2] and Brandom [1].

One might think of a sentence without occurrences of $\boldsymbol{T}$ as getting its semantic content not by inheriting it from another sentence, but in some more fundamental way. Such a sentence might be thought of *grounded*, since its content is grounded in the world of nonsemantic facts. Suppose that the sentence A is grounded, and that the name a refers to A. Then $\boldsymbol{T}a$ would also seem to be grounded, by virtue of inheriting its content from A. If b, however, refers to $\boldsymbol{T}b$ or to $\neg\boldsymbol{T}b$, then there is no way to find a grounded sentence for $\boldsymbol{T}b$ to inherit its content from. Thus $\boldsymbol{T}b$ is, on this informal analysis, ungrounded. Ungrounded sentences, it seems, cannot be either true or false: they do not have the right kind of content. There are tricky issues concerning composite sentences: if A is groundedly false and B is ungrounded, what is the status of $(A\ \&\ B)$? This looks very much like the time to select an evaluation scheme.

Grover [2] suggests that Kripke's [8] technical notion of groundedness is the best formal explication of the informal notion of groundedness, in terms of content-inheritance, that we have been articulating. Recall that a sentence is *grounded* for Kripke iff it gets the value **t** or **f** in the least fixed point. If we consider the construction of the least fixed point from the null hypothesis, then Grover's suggestion looks compelling. Fix a ground model $\mathrm{M} = \langle \mathrm{D}, \mathrm{I} \rangle$, and let ρ be some evaluation scheme. Let $\boldsymbol{S}$ be the ρ_{M}-sequence that builds $\mathrm{lfp}(\rho_{\mathrm{M}})$ up from the null hypothesis—i.e., $\boldsymbol{S}_0(\mathrm{d}) = \mathbf{n}$ for every $\mathrm{d} \in \mathrm{D}$. At the first stage of the revision process, every sentence with no occurrences of $\boldsymbol{T}$ gets a definite truth value. At each subsequent stage in the revision process, more and more sentences get definite truth-values: if $\boldsymbol{S}_\alpha(A) = \mathbf{t}$ or $\mathbf{f}$ and $\mathrm{I}(a) = A$, then $\boldsymbol{S}_{\alpha+1}(\boldsymbol{T}a) = \boldsymbol{S}_\alpha(A)$. Thus $\boldsymbol{T}a$ can be seen as inheriting its content from A.

Of course, we have to decide whether $(\boldsymbol{T}a\ \&\ \boldsymbol{T}b)$ gets a definite truth-value at stage $\alpha + 1$, when at stage α, $A = \mathrm{I}(a)$ is false and $B = \mathrm{I}(b)$ has not yet been assigned a definite truth-value. This depends on the evaluation scheme. The Kleene evaluation schemes seem the most intuitive, since the supervaluation schemes can evaluate a composite sentence as grounded even when none of its parts is grounded: for example, if the name b refers to the sentence $\neg\boldsymbol{T}b$, then on any of the supervaluation schemes, $(\boldsymbol{T}b \vee \neg\boldsymbol{T}b)$ is grounded though neither of its disjuncts is. The question might arise: where did ($\boldsymbol{T}b \vee \neg\boldsymbol{T}b$) get its content, if not from its disjuncts? Perhaps from its logical form? At any rate, if Grover is right, then $\mathbf{T}^{\mathrm{lfp},\, \rho}$ seems like a good theory of truth, where the evaluation scheme ρ is probably μ or κ.

Suppose that *L* is a language that contains no nonquote names, no function symbols and no nonlogical predicates. Suppose that L^+ is *L* enriched with the unary predicate ***T***, and that *L* has quote names for the sentences of L^+. And let M be any ground model for *L*. M displays no vicious reference, in our sense, on any of the theories of truth.

Consider the sentence $A = \forall x(\boldsymbol{T}x \vee \neg\boldsymbol{T}x)$. On the formal analysis of groundedness, *A* is ungrounded if we evaluate composite sentences using either μ or κ, despite the apparent absence of vicious reference. Is *A* ungrounded in our intuitive sense? If *A* is grounded, it must be true. So if we are using μ or κ, each instance of *A* must be true. In particular ($\boldsymbol{T}$‘A’ $\vee$ $\neg\boldsymbol{T}$‘A’) must be true. So, if we are using μ or κ, either ***T***‘*A*’ or ¬***T***‘*A*’ must be true. So if *A* is going to inherit its content, then *A* is going to have to inherit its content in part from itself. But this is, intuitively, sufficient for a sentence to be ungrounded. So, despite the absence of vicious reference, *A* seems ungrounded in our intuitive sense. We can imagine coming to this position in an attempt to formalize an anaphoric analysis of truth, independently of any concerns involving the liar’s paradox or any other paradox.

If non-vicious reference needs to be thought of in the way that we have defined it, then the story about grounding might trump any intuitions we have that blame truth’s nonclassical behaviour on vicious reference. One response is that there actually is vicious reference, because the quote name ‘*A*’ viciously refers to the ungrounded sentence *A*. But the above argument for the ungroundedness of $\forall x(\boldsymbol{T}x \vee \neg\boldsymbol{T}x)$ can be modified to work even when *L* has no quote names. In this case, if there is vicious reference anywhere, then it is in the bound variable *x*: such a variable can be thought of as referring indeterminately to all of the objects in the range of quantification. Among other things, *x* refers to the sentence $\forall x(\boldsymbol{T}x \vee \neg\boldsymbol{T}x)$ itself. On this line, since $\forall x(\boldsymbol{T}x \vee \neg\boldsymbol{T}x)$ is ungrounded no matter what the ground model is, there is *always* vicious reference in any language. But it is a kind of vicious reference that has no apparent relationship to the kind of vicious reference that has traditionally been seen as a source of paradox or pathology.

Acknowledgments. Many thanks to Anil Gupta, who responded to earlier drafts of this paper with very helpful criticisms and comments. He cleared up many confusions; without him, my misunderstanding of the issues at hand would even worse than it is now. Thanks also to Michael Kremer, for helpful technical and methodological remarks.

REFERENCES

[1] R. Brandom 1994, *Making it Explicit: Reasoning, Representing and Discursive Commitment*, Harvard University Press.

[2] D. Grover, "Inheritors and paradox", *Journal of Philosophy*, 590-604.

[3] D. Grover, J. Camp and N. Belnap 1975, "A prosentential theory of truth", *Philosophical Studies* 27, 73-125.

[4] A. Gupta 1982, "Truth and paradox", *Journal of Philosophical Logic* 11, 1-60. Reprinted in *Recent Essays on Truth and the Liar Paradox* (R.L. Martin, ed.), Oxford University Press, 1984, 175-236. Page references are to this reprinting.

[5] A. Gupta and N. Belnap 1993, *The Revision Theory of Truth,* MIT Press.

[6] M. Kremer 1988, "Kripke and the logic of truth", *Journal of Philosophical Logic*, 225-278.

[7] P. Kremer 2001, "Comparing fixed point and revision theories of truth", manuscript.

[8] S. Kripke 1975, "Outline of a theory of truth", *Journal of Philosophy*, 690-716.

[9] R.L. Martin and P.W. Woodruff 1975, "On representing ‘True-in-L’ in L", *Philosophia* 5, 217-221.

[10] V. McGee 1991, *Truth, Vagueness and Paradox: An Essay on the Logic of Truth*, Hackett, Indianapolis.

[11] S. Yablo 1993, "Paradox without self-reference", *Analysis* 53, 251-252.