J Philos Logic (2010) 39:345–367
DOI 10.1007/s10992-010-9136-4

# How Truth Behaves When There's No Vicious Reference

**Philip Kremer**

Received: 17 June 2009 / Accepted: 13 August 2009 / Published online: 19 June 2010


**Abstract** In *The Revision Theory of Truth* (MIT Press), Gupta and Belnap (1993) claim as an advantage of their approach to truth "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." To clarify this remark, they define *Thomason* models, nonpathological models in which truth behaves like a classical concept, and investigate conditions under which a model is Thomason: they argue that a model is Thomason when there is no vicious reference in it. 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.

**Keywords** Truth · Paradox · Vicious reference · Fixed-point semantics · Revision theory

## 1 Introduction

Two factors seem to be at play in the truth-theoretic paradoxes: intuitive principles concerning truth; and the facts about which singular terms refer to which sentences, and so on. For example, paradoxicality might be partially

P. Kremer (✉)
Department of Philosophy, University of Toronto, Toronto, ON M5R 2M8, Canada
e-mail: philip.kremer@utoronto.ca

attributed to the contingent fact that the singular term, "the italicized sentence on page 2", refers to the sentence,

*The italicized sentence on page 2 is not true.*

Factors of the second kind might be represented by a ground model: an interpretation of all the names, function symbols, and relation symbols in the potentially self-referential language under study, with the exception of the predicate "$x$ is true". Formally, suppose that $\mathcal{L}$ is an uninterpreted first order language. $M = \langle D, I\rangle$ is a *classical model for* $\mathcal{L}$ iff $D$ is a nonempty set and $I$ is a function assigning to each name of $\mathcal{L}$ a member of $D$, to each $n$-place function symbol of $\mathcal{L}$ an $n$-place function on $D$, and to each $n$-place relation symbol a function from $D^n$ to $\{\mathbf{t}, \mathbf{f}\}$. Suppose that $\mathcal{L}$ and $\mathcal{L}^+$ are first-order languages, where $\mathcal{L}^+$ is $\mathcal{L}$ expanded with a distinguished predicate (one-place relation symbol) $\boldsymbol{T}$, and where $\mathcal{L}$ has a quote name '$A$' for each sentence $A$ of $\mathcal{L}^+$. Then $\mathcal{L}$ and $\mathcal{L}^+$ are a corresponding *ground language* and *truth language*. We follow Gupta and Belnap [6] in defining $S =_{\text{df}} \{A : A \text{ is a sentence of } \mathcal{L}^+\}$. A *ground model for* $\mathcal{L}$ is a classical model $M = \langle D, I\rangle$ for $\mathcal{L}$ such that $I(\text{'}A\text{'}) = A \in D$ for each $A \in S$. A ground model tells us what the terms in $\mathcal{L}^+$ refer to and what the extensions are of the nonsemantic predicates.

We might want to expand a ground model $M = \langle D, I\rangle$ for $\mathcal{L}$ to a classical model $M' = \langle D, I'\rangle$ for $\mathcal{L}^+$ so that the extension[1] of $\boldsymbol{T}$ in $M'$ is the set of sentences of $\mathcal{L}^+$ true in $M'$: we will call such a model $M'$ a *Tarski* model. A Tarski model for $\mathcal{L}^+$ is one in which truth behaves as it intuitively should. Unfortunately, some ground models cannot be expanded to Tarski models. Suppose that the ground language $\mathcal{L}$ has one nonquote name, $b$, no function symbols, and no relation symbols. Also suppose that $M = \langle D, I\rangle$ is a ground model, with $D = S$ and $I(b) = \neg\boldsymbol{T}b$. Finally suppose that $M' = \langle D, I'\rangle$ is a classical model expanding $M$. Note that $I'(\boldsymbol{T})(\neg\boldsymbol{T}b) = \mathbf{t}$ iff $I'(\boldsymbol{T})(I'(b)) = \mathbf{t}$ 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 expand $M$ to a Tarski model—and to the fact that the name $b$ refers to the sentence $\neg\boldsymbol{T}b$.

Suppose that $I(b) = \boldsymbol{T}b$ rather than $\neg\boldsymbol{T}b$. Applying ordinary reasoning about truth to this *truth-teller* does not lead to contradiction, as it does with the liar: in this case, we *can* expand $M$ to a Tarski model. Indeed, we can expand $M$ to *two* Tarski models, $M' = \langle D, I'\rangle$ and $M'' = \langle D, I''\rangle$, where $I'(\boldsymbol{T})(\boldsymbol{T}b) = \mathbf{t}$ and $I''(\boldsymbol{T})(\boldsymbol{T}b) = \mathbf{f}$. The problem with the truth-teller is that there seems to be nothing to decide between $M'$ and $M''$: so the truth-teller is still pathological if not paradoxical.[2] An example in [12] raises the question whether self-reference

[1] The *extension* of an $n$-place relation symbol $R$ is $\{\overline{d} \in D^n : I(R)(\overline{d}) = \mathbf{t}\}$.

[2] Anil Gupta commended this distinction and terminology to me, in correspondence.

or circular reference is necessary for paradoxicality.[3] But even in noncircular examples the pathology might be attributed to vicious reference in the ground model. Finally, vicious reference need not involve singular terms. Consider a ground language $\mathcal{L}$ with a one-place predicate $G$, and a ground model $M = \langle D, I\rangle$ where $D = S$ and $I(G)(A) = \mathbf{t}$ iff $A = \forall x(Gx \supset \neg \boldsymbol{T}x)$. $M$ cannot be expanded to a Tarski model. Here it is a predicate, $G$, that is viciously referring.

Not all self-reference is vicious or pathology-producing. Consider,

*The italicized sentence on page 3 contains four words.*

This is an unpathologically false self-referential sentence. Formally, suppose that the ground language $\mathcal{L}$ has exactly one nonquote name, $c$, one one-place 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 there is nothing vicious about the self-reference: $Gc$ is unpathologically false, and reference to it is non-vicious.

The two notions of a *pathological* ground model and of a ground model with *vicious reference* suggest the complementary notions of a *nonpathological* ground model and of a ground model with *no vicious reference*. Gupta and Belnap [6] claim that one advantage of their 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,[4] the ground models in which—from Gupta and Belnap's particular theoretic perspective—truth behaves like a classical concept.[5] They investigate some of the conditions under which a ground model is Thomason and 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.[6]

In the current paper, we approach the topic from a perspective slightly different from Gupta and Belnap's, but prompted by their discussion. We consider a number of theories of truth, both Gupta and Belnap's revision theories and theories motivated by the fixed-point semantics[7] in opposition to which Gupta and Belnap develop their approach. We give a formal definition of when there is no vicious reference in a ground model relative to this or that

[3]The example is of an infinite sequence of sentence $A_1, A_2, \ldots$, where $A_n$ states that, for every $m > n$, $A_m$ is not true. Whether this example is circular is discussed in [2].

[4]See Definition 3.1, below.

[5]They 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.

[6]Gupta emphasized this in correspondence.

[7]See Section 2.1, below.

theory **T**, and a formal definition of when truth behaves like a classical concept in a ground model relative to this or that theory **T**. We 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 [5] 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 assertible in the model $M$." (p. 194) We will show that some of the fixed-point theories satisfy our desideratum, and that at least one of 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 to 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 is theory-relative. We will wait until our formal definitions are on the table before discussing these issues.

## 2 Fixed-point and Revision Theories of Truth

Sections 2 and 3 of Kremer [8] develop the fixed-point semantics [9, 10] and the revision theoretic semantics [6] for languages expressing their own truth concepts. Here, we repeat the main definitions without the discussion and motivation. We will define ten fixed-point theories of truth and three revision theories. As pointed out in [8], each of these thirteen theories relies on what M. Kremer [7] calls the *supervenience of semantics*: the intuition that the interpretation of $\boldsymbol{T}$ should be determined by the interpretation of the nonsemantic names, function symbols and relation symbols, as represented by a ground model. M. Kremer argues both that that Kripke [9] does not endorse this proposal, and that this proposal misinterprets the fixed-point semantics. We consider the desideratum—that if there is no vicious reference, then truth should behave like a classical concept—in a nonsupervenience setting in Appendix B.

### 2.1 Fixed-point Theories

A *three-valued model* for a first-order language $\mathcal{L}$ is just like a classical model, except that the function $I$ assigns, to each $n$-place relation symbol, a function

from $D^n$ to $\{\mathbf{t}, \mathbf{f}, \mathbf{n}\}$. A classical model is a special case of a three-valued model. Officially **t**(rue), **f**(alse) and **n**(either) are three truth values, but **n** can be thought of as the absence of a truth value. We order the truth values as follows: $\mathbf{n} \leq \mathbf{n} \leq \mathbf{t} \leq \mathbf{t}$ and $\mathbf{n} \leq \mathbf{n} \leq \mathbf{f} \leq \mathbf{f}$. We say that $M = \langle D, I\rangle \leq M' = \langle D, I'\rangle$ iff $I(X) = I'(X)$ for each name or function symbol $X$, and $I(R)(d_1, \ldots, d_n) \leq I'(R)(d_1, \ldots, d_n)$ for each $n$-place relation symbol $R$ and each $d_1, \ldots, d_n \in D$. Given a three-valued model $M = \langle D, I\rangle$ and an assignment $s$ of values to the variables, the value $Val_{M,s}(t) \in D$ of each term $t$ is defined in the standard way. The atomic formula $Rt_1 \ldots t_n$ is assigned the truth value $I(R)(Val_{M,s}(t_1), \ldots, Val_{M,s}(t_n))$. To evaluate composite formulas, we must have some *evaluation scheme*: for example, if $A$ is **f**(alse) and $B$ is **n**(either), then we must decide whether $(A \,\&\, B)$ is **f** or **n**.

For classical models, we use the standard classical evaluation scheme, $\tau$: If $M$ is a classical model for $\mathcal{L}$ and $A$ is a sentence of $\mathcal{L}$, then $Val_{M,\tau}(A)$ is the standard truth value of $A$ in $M$. For nonclassical three-valued models, we consider the *weak Kleene scheme*, $\mu$, and the *strong Kleene scheme*, $\kappa$. The Kleene schemes treat negation identically: $\neg\mathbf{t} = \mathbf{f}$, $\neg\mathbf{f} = \mathbf{t}$, and $\neg\mathbf{n} = \mathbf{n}$. They differ in their treatment of conjunction as in the following truth table:

| $A$ | $B$ | $A \,\&\, B$, with $\mu$ | $A \,\&\, B$, with $\kappa$ |
|---|---|---|---|
| **t** | **t** | **t** | **t** |
| **t** | **f** | **f** | **f** |
| **t** | **n** | **n** | **n** |
| **f** | **t** | **f** | **f** |
| **f** | **f** | **f** | **f** |
| **f** | **n** | **n** | **f** |
| **n** | **t** | **n** | **n** |
| **n** | **f** | **n** | **f** |
| **n** | **n** | **n** | **n** |

If we treat universal quantification analogously to conjunction, then for each sentence $A$ and for the weak and strong Kleene schemes, $\mu$ and $\kappa$, we can define $Val_{M,\mu}(A)$ and $Val_{M,\kappa}(A)$: the truth value of $A$ in $M$ according to $\mu$ and the truth value of $A$ in $M$ according to $\kappa$.

We also consider van Fraassen's *supervaluation* scheme, $\sigma$:

$$Val_{M,\sigma}(A) = \begin{cases} \mathbf{t}, & \text{if } Val_{M',\tau}(A) = \mathbf{t} \text{ for every classical } M' \geq M \\ \mathbf{f}, & \text{if } Val_{M',\tau}(A) = \mathbf{f} \text{ for every classical } M' \geq M \\ \mathbf{n}, & \text{otherwise.} \end{cases}$$

Suppose that $\mathcal{L}$ and $\mathcal{L}^+$ are a corresponding ground language and truth language, and that $M = \langle D, I\rangle$ is a classical ground model for $\mathcal{L}$. An *hypothesis* is a function $h : D \to \{\mathbf{t}, \mathbf{f}, \mathbf{n}\}$, and a *classical* hypothesis, a function $h : D \to \{\mathbf{t}, \mathbf{f}\}$. Say that $h \leq h'$ iff $h(d) \leq h'(d)$ for every $d \in D$. A function $\mathcal{F}$ on hypotheses is *monotone* iff, for all hypotheses $h$ and $h'$, if $h \leq h'$ then $\mathcal{F}(h) \leq \mathcal{F}(h')$. Let $M + h$ be the model $M' = \langle D, I'\rangle$ for $\mathcal{L}^+$, where $I'$ and $I$ agree on the constants of $\mathcal{L}$ and where $I'(\boldsymbol{T}) = h$. Models of the form $M + h$ are *expanded* models.

For $\rho = \tau, \mu, \kappa$, or $\sigma$, define the *jump operator* $\rho_M$ on the set of hypotheses as follows, restricting the definition to classical hypotheses where $\rho = \tau$:

$$\rho_M(h)(A) = Val_{M+h,\rho}(A), \text{ if } A \in S = \text{ the set of sentences of } \mathcal{L}^+$$
$$\rho_M(h)(d) = \mathbf{f}, \text{ if } d \in D - S.$$

Note that $\mu_M$, $\kappa_M$ and $\sigma_M$ are monotone, for every ground model $M$.

**Theorem 2.1** (Kripke [9]) *Each total monotone function $\mathcal{F}$ on hypotheses has a least fixed point, $lfp(\mathcal{F})$.*

Hypotheses $h$ and $h'$ are *compatible* iff $h \leq h''$ and $h' \leq h''$ for some hypothesis $h''$; and $h$ is *$\mathcal{F}$-intrinsic* iff $h$ is compatible with every fixed point of $\mathcal{F}$.

**Theorem 2.2** (Kripke [9]) *Each total monotone function $\mathcal{F}$ has a greatest intrinsic fixed point, $gifp(\mathcal{F})$, which is not in general identical to $lfp(\mathcal{F})$.*

**Definition 2.3** Let $\rho = \mu$, $\kappa$ or $\sigma$. The sentence $A$ of $\mathcal{L}^+$ is *valid in the ground model $M$ according to* (the theory) $\mathbf{T}^{lfp,\rho}$ [$\mathbf{T}^{gifp,\rho}$] iff $lfp(\rho_M)(A) = \mathbf{t}$ [$gifp(\rho_M)(A) = \mathbf{t}$].

A hypothesis $h$ is *weakly consistent* iff the set of sentences $\{A \in S : h(A) = \mathbf{t}\}$ is consistent, and *strongly consistent* iff $\{A \in S : h(A) = \mathbf{t}\} \cup \{\neg A : A \in S \ \&\ h(A) = \mathbf{f}\}$ is consistent. The jump operators $\sigma 1$ and $\sigma 2$ are defined for weakly and strongly consistent hypotheses, respectively, as follows:

$$\sigma 1_M(h)(A) = \begin{cases} \mathbf{t}, \text{ if } \tau_M(h')(A) = \mathbf{t} \text{ for every weakly consistent classical } h' \geq h \\ \mathbf{f}, \text{ if } \tau_M(h')(A) = \mathbf{f} \text{ for every weakly consistent classical } h' \geq h \\ \mathbf{n}, \text{ otherwise, for sentences } A \in S. \end{cases}$$

$$\sigma 1_M(h)(d) = \mathbf{n}, \text{ for } d \in D - S.$$

$$\sigma 2_M(h)(A) = \begin{cases} \mathbf{t}, \text{ if } \tau_M(h')(A) = \mathbf{t} \text{ for every strongly consistent classical } h' \geq h \\ \mathbf{f}, \text{ if } \tau_M(h')(A) = \mathbf{f} \text{ for every strongly consistent classical } h' \geq h \\ \mathbf{n}, \text{ otherwise, for sentences } A \in S. \end{cases}$$

$$\sigma 2_M(h)(d) = \mathbf{n}, \text{ for } d \in D - S.$$

The operator $\sigma 1_M$ [$\sigma 2_M$] is monotone on the weakly [strongly] consistent hypotheses. This suffices for $\sigma 1_M$ [$\sigma 2_M$] to have both a least fixed point and a greatest intrinsic fixed point. We treat $\sigma 1$ and $\sigma 2$ as two new evaluation schemes. Theories $\mathbf{T}^{lfp,\sigma 1}$, $\mathbf{T}^{gifp,\sigma 1}$, $\mathbf{T}^{lfp,\sigma 2}$, and $\mathbf{T}^{gifp,\sigma 2}$ are defined as in Definition 2.3, above.

Kripke [9] uses the least fixed point and the greatest intrinsic fixed point to define certain properties of sentences. Fix an evaluation scheme $\rho$, a ground

model $M = \langle D, I\rangle$ for $\mathcal{L}$, and a sentence $A$ of $\mathcal{L}^+$. We say that $A$ is *$\rho$-grounded in $M$* iff $lfp(\rho_M)(A) \neq \mathbf{n}$, and *$\rho$-intrinsic in $M$* iff $gifp(\rho_M)(A) \neq \mathbf{n}$.

### 2.2 Revision Theories

Given any function $\mathcal{F}$ on hypotheses, an *$\mathcal{F}$-sequence*, or a *revision sequence for $\mathcal{F}$*, is an ordinal-length sequence $\mathscr{S}$ of hypotheses such that $\mathscr{S}_{\alpha+1} = \mathcal{F}(\mathscr{S}_\alpha)$, for every ordinal $\alpha$; and every limit ordinal $\lambda$, every truth value $\mathbf{x}$, and every $d \in D$, we have

$$\mathscr{S}_\lambda(d) = \mathbf{x} \quad \text{if there is a } \beta < \lambda \text{ such that } \mathscr{S}_\alpha(d) = \mathbf{x} \text{ for every ordinal } \alpha \text{ with } \beta \leq \alpha < \lambda.$$

Note that if $\mathscr{S}$ is an $\mathcal{F}$-sequence, then $\mathcal{F}$ is defined on $\mathscr{S}_\alpha$ for every ordinal $\alpha$; so, if $\mathscr{S}$ is a $\tau_M$-sequence, then $\mathscr{S}_\alpha$ is classical for every ordinal $\alpha$. Any ordinal-length sequence $\mathscr{S}$ of hypotheses *culminates* in $h$ iff there is an ordinal $\beta$ such that $\mathscr{S}_\alpha = h$ for every ordinal $\alpha \geq \beta$. Note that if $\rho = \mu, \kappa, \sigma, \sigma 1$, or $\sigma 2$ and if $M$ is a ground model, then there is a unique $\rho_M$-sequence $\mathscr{S}$ such that $\mathscr{S}_0(d) = \mathbf{n}$ for every $d \in D$. Furthermore, $\mathscr{S}$ culminates in $lfp(\rho_M)$.

A sentence $A$ of $\mathcal{L}^+$ is *stably* **t** [**f**] in the $\tau_M$-sequence $\mathscr{S}$ iff there is an ordinal $\beta$ such that for every $\gamma \geq \beta$, we have $\mathscr{S}_\gamma(A) = \mathbf{t}$ [**f**]; and *nearly stably* **t** [**f**] in $\mathscr{S}$ iff there is an ordinal $\beta$ such that for every $\gamma \geq \beta$, there is a natural number $m$ such that for every $n \geq m$, we have $\mathscr{S}_{\gamma+n}(A) = \mathbf{t}$ [**f**]. A $\tau_M$-sequence $\mathscr{S}$ is *maximally consistent* iff $\mathscr{S}_\alpha$ is strongly consistent for every ordinal $\alpha$.

**Definition 2.4** Suppose that $M$ is a ground model for the ground language $\mathcal{L}$. The sentence $A$ of $\mathcal{L}^+$ is *valid in $M$ according to* (the theory) $\mathbf{T}^*$ [$\mathbf{T}^\#$, $\mathbf{T}^c$] iff $A$ is stably **t** in every $\tau_M$-sequence [nearly stably **t** in every $\tau_M$-sequence, stably **t** in every maximally consistent $\tau_M$-sequence].

**Definition 2.5** Suppose that $M$ is a ground model for the ground language $\mathcal{L}$, and that $\mathbf{T}$ is one of the thirteen theories of truth under consideration. $\mathbf{V}^{\mathbf{T}}_M =_{\text{df}} \{A \in S : A \text{ is valid in } M \text{ according to } \mathbf{T}\}$. And $\neg\mathbf{V}^{\mathbf{T}}_M =_{\text{df}} \left\{A \in S : \neg A \in \mathbf{V}^{\mathbf{T}}_M\right\}$.

## 3 Truth Behaving like a Classical Concept

Consider a classical ground model $M = \langle D, I\rangle$ that makes no distinctions, other than with quote names, among the sentences of $\mathcal{L}^+$: for an extreme case, suppose that $\mathcal{L}$ has no nonquote names, no function symbols and no nonlogical relation symbols. There seems to be no opportunity for vicious reference under these circumstances. And yet $lfp(\mu_M)$ and $lfp(\kappa_M)$ are nonclassical: thus it seems that neither of the least-fixed-point theories $\mathbf{T}^{lfp,\mu}$ or $\mathbf{T}^{lfp,\kappa}$ dictates that truth behaves like a classical concept in $M$. This is a simple example of what 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 $lfp(\sigma_M)$ *is* classical in $M$, by Corollary 4.26 in [8].)

Gupta and Belnap introduce their notion of a *Thomason* model (see [8], 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 3.1** A ground model $M$ is *Thomason* iff all $\tau_M$-sequences culminate in one and the same fixed point.

The notion of a Thomason model is a formalization 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 some nonclassical jump operator rather than in terms of $\tau_M$-sequences.

Soon after introducing Thomason models, Gupta and Belnap note that for every Thomason model $M$, the classical jump operator $\tau_M$ has a unique fixed point, say $h$.[8] They point out that, if $M$ is Thomason, then $A \in \mathbf{V}^{\mathbf{T}^*}_M$ iff $A \in \mathbf{V}^{\mathbf{T}^\#}_M$ iff $h(A) = \mathbf{t}$ for every sentence $A$ of $\mathcal{L}^+$. 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 [8].

**Definition 3.2** Suppose that $\mathbf{T}$ is one of our thirteen theories of truth. $\mathbf{T}$ *dictates that truth behaves like a classical concept in the ground model* $M$ iff $\mathbf{V}^{\mathbf{T}}_M \cup \neg\mathbf{V}^{\mathbf{T}}_M = S$.

Note that the notion of a Thomason model is a notational variant of of the notion of a ground model in which $\mathbf{T}^*$ dictates that truth behaves like a classical concept.

## 4 No Vicious Reference

Gupta and Belnap never give a precise definition of *no vicious reference*, but they do suggest a way to proceed. Gupta [5] 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 $\mathcal{L}^+$: "Now, what sorts of self-reference can we allow in $\mathcal{L}$ [interpreted via $M$]? What kinds of distinctions among the

[8] $\tau_M$ having a unique fixed point is a necessary but not a sufficient condition for $M$ to be Thomason.

sentences containing the truth predicate can we make without violating the fundamental intuition?"[9] (p. 191) This second question is made more precise with Gupta and Belnap's notion of a name's, relation symbol's, or function symbol's interpretation being neutral relative to some subset $X$ of the domain of discourse $D$.

**Definition 4.1** ([6], Definition 2D.2) Suppose that $M = \langle D, I \rangle$ is a ground model for the ground language $\mathcal{L}$ and that $X \subseteq D$.

1. The interpretation of a name $c$ is *$X$-neutral* iff $I(c) \notin X$.
2. The interpretation of an $n$-place relation symbol $R$ is *$X$-neutral* iff for all $d_1, \ldots, d_n, d_i' \in D$, if $d_i, d_i' \in X$ then $I(R)(d_1, \ldots, d_i, \ldots, d_n) = I(R)(d_1, \ldots, d_i', \ldots, d_n)$.
3. The interpretation of an $n$-place function symbol $f$ is *$X$-neutral* iff both
   a. the range of $I(f)$ is disjoint from $X$; and
   b. for all $d_1, \ldots, d_n, d_i' \in D$, if $d_i, d_i' \in X$ then $I(f)(d_1, \ldots, d_i, \ldots, d_n) = I(f)(d_1, \ldots, d_i', \ldots, d_n)$.

**Definition 4.2** ([6], Definition 6A.2) A ground model $M = \langle D, I \rangle$ is *$X$-neutral* iff the interpretations in $M$ of all the nonquote names, nonlogical relation symbols, and function symbols are $X$-neutral.

## 4.1 No Vicious Reference: A Crescendo of Results

Gupta and Belnap prove a crescendo of results relating a ground model's ability to make distinctions among the sentences of $\mathcal{L}^+$ to its Thomasonness–i.e., to whether $\mathbf{T}^*$ dictates that truth behaves like a classical concept in it. Here goes:[10]

**Theorem 4.3** *If $M$ is S-neutral, then $M$ is Thomason.*

**Theorem 4.4** *If $M$ is $X$-neutral where $X \subseteq D$ contains all the sentences that have occurrences of $\boldsymbol{T}$ then $M$ is Thomason.*

**Theorem 4.5** *If $M$ is $X$-neutral where $X \subseteq D$ contains all the $\mu$-ungrounded sentences then $M$ is Thomason.*

**Theorem 4.6** *If $M$ is $X$-neutral where $X \subseteq D$ contains all the $\kappa$-ungrounded sentences then $M$ is Thomason.*

[9] 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." ([5], p. 181).

[10] See [6], Theorems 6A.5 and 6B.4.

**Theorem 4.7** *If $M$ is $X$-neutral where $X \subseteq D$ contains all the $\sigma$-ungrounded sentences then $M$ is Thomason.*

These results are part of Gupta and Belnap's 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 $\mathcal{L}^+$—then $M$ is Thomason. As already noted, the Thomason ground models are precisely the ground models in which $\mathbf{T}^*$ dictates that truth behaves like a classical concept. Thus, Gupta and Belnap's argument is a case-by-case argument that $\mathbf{T}^*$ satisfies the following Gupta–Belnap Desideratum on theories $\mathbf{T}$ of truth:

*Gupta–Belnap Desideratum*[11] (GBD) If there is no vicious reference in the ground model $M$ then $\mathbf{T}$ dictates that truth behaves like a classical concept in $M$.

As mentioned in Section 1, both the GBD and the argument that $\mathbf{T}^*$ satisfies it rely on an informal, intuitive, theory-neutral notion of non-vicious reference. The general strategy implicit in the argument is as follows: (1) find some set $Y$ of intuitively unproblematic sentences; (2) show that if $M$ is $(S - Y)$-neutral—if $M$ cannot make any distinctions among potentially problematic sentences—then $M$ is Thomason. In Theorem 4.3, $Y = \emptyset$. In Theorem 4.4, $Y$ is any set of $\boldsymbol{T}$-free sentences. In Theorem 4.5, $Y$ is any set of $\mu$-grounded sentences. In Theorem 4.6 [4.7], $Y$ is any set of $\kappa$-grounded [$\sigma$-grounded] sentences. There are fairly strong theory-neutral intuitions that, in each of these cases, reference 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 GBD (rather, only a counterexample to the claim that $\mathbf{T}^*$ satisfies a converse desideratum).

### 4.2 No Vicious Reference: Extending the Gupta–Belnap Argument

There are two natural ways to extend Gupta and Belnap's informal considerations:

1. Extend Theorems 4.3–4.7 to other sets of sentences.
2. Test whether analogues to Theorems 4.3–4.7 apply to other theories of truth, both fixed-point and revision.

Pursuing (1), we consider two conjectures:

**Conjecture 4.8** If $M$ is $X$-neutral where $X \subseteq D$ contains all the $\sigma 1$-ungrounded sentences then $M$ is Thomason.

**Conjecture 4.9** If $M$ is $X$-neutral where $X \subseteq D$ contains all the $\sigma 2$-ungrounded sentences then $M$ is Thomason.

[11] This desideratum is not quite explicit in either [6] or [5], but we take it to be implicit.

Each of these conjectures is false, as follows from Theorem 4.21 of [8]. Already, this spells trouble for Gupta and Belnap's case-by-case argument: their conclusion would require reference to some $\sigma 1$-grounded or some $\sigma 2$-grounded sentences to be vicious.

Pursuing (2), for each theory **T** of truth, we can consider an analogue of each of Theorems 4.3–4.7:

**Conjecture 4.10** (Analogous to Theorem 4.3) *If M is S-neutral then* **T** *dictates that truth behaves like a classical concept in M.*

**Conjecture 4.11** (Analogous to Theorem 4.4) *If M is X-neutral where* $X \subseteq D$ *contains all the sentences that have occurrences of* ***T*** *then* **T** *dictates that truth behaves like a classical concept in M.*

**Conjecture 4.12** (Analogous to Theorem 4.5) *If M is X-neutral where* $X \subseteq D$ *contains all the* $\mu$*-ungrounded sentences then* **T** *dictates that truth behaves like a classical concept in M.*

**Conjecture 4.13** (Analogous to Theorem 4.6) *If M is X-neutral where* $X \subseteq D$ *contains all the* $\kappa$*-ungrounded sentences then* **T** *dictates that truth behaves like a classical concept in M.*

**Conjecture 4.14** (Analogous to Theorem 4.7) *If M is X-neutral where* $X \subseteq D$ *contains all the* $\sigma$*-ungrounded sentences then* **T** *dictates that truth behaves like a classical concept in M.*

None of these conjectures holds for either $\mathbf{T}^{lfp,\mu}$ or $\mathbf{T}^{lfp,\kappa}$, since $lfp(\mu_M)$ and $lfp(\kappa_M)$ are never classical. So $\mathbf{T}^{lfp,\mu}$ and $\mathbf{T}^{lfp,\kappa}$ fail to satisfy the GBD. Conjectures 4.10–4.13 hold for the remaining eleven theories among our thirteen theories. So our case-by-case considerations so far suggest that all eleven of these theories satisfy the GBD. Conjecture 4.14 fails for $\mathbf{T}^{lfp,\mu}$, $\mathbf{T}^{lfp,\kappa}$, and $\mathbf{T}^{lfp,\sigma}$ and holds for the remaining ten of our thirteen theories. So our case-by-case considerations so far only rule out $\mathbf{T}^{lfp,\mu}$, $\mathbf{T}^{lfp,\kappa}$, and $\mathbf{T}^{lfp,\sigma}$ as satisfying the GBD.[12]

## 4.3 No Vicious Reference: Gupta and Belnap's Last Step

The last step in Gupta and Belnap's argument is the following:

**Theorem 4.15** ([6], Theorem 6B.8) *Suppose that M is an* $(S - Y)$*-neutral model and that Y contains only sentences that are either stably* **t** *in all* $\tau_M$*-sequences or stably* **f** *in all such sequences—in other words,* $Y \subseteq \mathbf{V}_M^{\mathbf{T}^*} \cup \neg\mathbf{V}_M^{\mathbf{T}^*}$*. Then, M is Thomason.*

[12] The claims in this paragraph follow from Theorem 4.21 of [8].

Given Theorem 4.21 in [8], we can strengthen Theorem 4.15 as follows:

**Theorem 4.16** *Suppose that $M$ is an $(S-Y)$-neutral model and that $Y \subseteq \mathbf{V}_M^{\mathbf{T}^*} \cup \neg\mathbf{V}_M^{\mathbf{T}^*}$. Then* $\mathbf{T}$ *dictates that truth behaves like a classical concept in $M$, where* $\mathbf{T} = \mathbf{T}^*$, $\mathbf{T}^{\#}$, $\mathbf{T}^{c}$, $\mathbf{T}^{gifp,\mu}$, $\mathbf{T}^{gifp,\kappa}$, $\mathbf{T}^{gifp,\sigma}$, $\mathbf{T}^{gifp,\sigma 1}$, *or* $\mathbf{T}^{gifp,\sigma 2}$.

Theorem 4.16 fails for $\mathbf{T} = \mathbf{T}^{lfp,\sigma 1}$ or $\mathbf{T}^{lfp,\sigma 2}$: see Example 6B.13 in [6]. Should we conclude that $\mathbf{T}^{lfp,\sigma 1}$ and $\mathbf{T}^{lfp,\sigma 2}$ fail to satisfy the GBD?

The conditions on the ground model in Theorems 4.3–4.7 might be clear cases of non-vicious reference, from an intuitive theory-neutral perspective. But the condition on the ground model $M$ in Theorems 4.15 and 4.16 is that $M$ be $(S-Y)$-neutral, where $Y \subseteq \mathbf{V}_M^{\mathbf{T}^*} \cup \neg\mathbf{V}_M^{\mathbf{T}^*}$. Thus, on the strategy we suggested for interpreting Theorems 4.3–4.7, the set $Y$ of unproblematic sentences can be any subset of $\mathbf{V}_M^{\mathbf{T}^*} \cup \neg\mathbf{V}_M^{\mathbf{T}^*}$. At this point in their case-by-case argument, Gupta and Belnap are no longer working with clearly theory-neutral intuitions concerning non-vicious reference: the intuitions at work are intuitions that rate as non-vicious any reference to the sentences that are stably true in all $\tau_M$-sequences or stably false in all $\tau_M$-sequences. This seems motivated by the revision theoretic semantics, and more specifically by the theory $\mathbf{T}^*$.

As noted, Theorem 4.16 fails for $\mathbf{T} = \mathbf{T}^{lfp,\sigma 1}$ or $\mathbf{T}^{lfp,\sigma 2}$. If we appeal to this as as evidence that $\mathbf{T}^{lfp,\sigma 1}$ and $\mathbf{T}^{lfp,\sigma 2}$ do not satisfy the GBD, 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 one-place predicate. Further, suppose that $I(c) = \boldsymbol{T}\text{`}Gc\text{'}$. The sentence $\boldsymbol{T}\text{`}Gc\text{'}$ does not have any truth-value in the ground model, since $I$ assigns no extension or anti-extension to $\boldsymbol{T}$. But on any reasonable theory of truth, $\boldsymbol{T}\text{`}Gc\text{'}$ will behave classically, and will be assigned the same classical truth-value as $Gc$. So reference to $\boldsymbol{T}\text{`}Gc\text{'}$ ought to be non-vicious.

Here, we already have an implicit relativization of non-vicious reference to the theory of truth: reference to $\boldsymbol{T}\text{`}Gc\text{'}$ should be non-vicious on any reasonable theory of truth since $\boldsymbol{T}\text{`}Gc\text{'}$ should behave classically on any reasonable theory of truth. Consider an unreasonable theory of truth, $\mathbf{T}_{\text{null}}$, which assigns to every sentence of the form $\boldsymbol{T}b$ the truth-value $\mathbf{n}$. According to $\mathbf{T}_{\text{null}}$, it is not so clear whether reference to $\boldsymbol{T}\text{`}Gc\text{'}$ is vicious: after all, according to $\mathbf{T}_{\text{null}}$, 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 on reasonable theories of truth. Thus even our

informal intuitions are, in some sense, theory-relative: relative to reasonable theories.

For the kind of argument Gupta and Belnap advance, 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 function symbols and relation symbols, non-vicious distinctions will be distinctions among the non-sentences together with the unproblematic sentences.

**Definition 4.17** Let $\mathbf{T}$ be any of our thirteen theories of truth. Let $M$ be a ground model for a ground language $\mathcal{L}$. $\mathbf{T}$ *dictates that there is no vicious reference in* $M$ iff $M$ is $(S - Y)$-neutral for some $Y \subseteq \mathbf{V}^{\mathbf{T}}_M \cup \neg\mathbf{V}^{\mathbf{T}}_M$.

## 5 How Truth Behaves When There's No Vicious Reference

The modified Gupta–Belnap desideratum is as follows:

*Modified Gupta–Belnap Desideratum* (MGBD) 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 $M$.

Our main theorem follows from Theorem 4.21, (2) and (3), in [8]:

**Theorem 5.1**

(1) $\mathbf{T}^*$, $\mathbf{T}^c$, $\mathbf{T}^{lfp,\sigma 2}$, $\mathbf{T}^{gifp,\mu}$, $\mathbf{T}^{gifp,\kappa}$, $\mathbf{T}^{gifp,\sigma}$, $\mathbf{T}^{gifp,\sigma 1}$ *and* $\mathbf{T}^{gifp,\mu}$ *satisfy the MGBD.*
(2) $\mathbf{T}^{\#}$, $\mathbf{T}^{lfp,\mu}$, $\mathbf{T}^{lfp,\kappa}$, $\mathbf{T}^{lfp,\sigma}$ *and* $\mathbf{T}^{lfp,\sigma 1}$ *do not.*

Thus, a number of the fixed-point theories satisfy the MGBD while Gupta and Belnap's revision theory $\mathbf{T}^{\#}$ does not.

We recall Gupta's caution (Section 1): the modified Gupta–Belnap desideratum 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 5.1 brings out a striking difference between the MGBD and the GBD. Say that $\mathbf{T}' \geq_1 \mathbf{T}$ iff for every language $\mathcal{L}$ every ground model $M$ and every sentence $A$ of $\mathcal{L}^+$, if $A$ is valid in $M$ according to $\mathbf{T}$ then $A$ is valid in $M$ according to $\mathbf{T}'$. (See Definition 4.1 in [8].) As Gupta has noted in correspondence, if a theory $\mathbf{T}$ satisfies the GBD then any theory $\mathbf{T}' \geq_1 \mathbf{T}$ is also bound to satisfy it. Not so with MGBD, which is satisfied by $\mathbf{T}^*$ but not by $\mathbf{T}^{\#}$, although $\mathbf{T}^{\#} \geq_1 \mathbf{T}^*$. The following example shows that $\mathbf{T}^{\#}$ does not satisfy the MGBD.

*Example 5.2* ([6], Example 6B.9, and [8], Example 5.7) Consider a ground language $\mathcal{L}$ with a one-place predicate $G$, and no other nonlogical vocabulary besides quote names. Let

$$\begin{array}{lcl} A & = & \exists x(Gx \,\&\, \neg \boldsymbol{T}x) \\ B & = & \exists x \exists y(Gx \,\&\, Gy \,\&\, \neg \boldsymbol{T}x \,\&\, \neg \boldsymbol{T}y \,\&\, x \neq y), \text{ and} \\ Y & = & \{\boldsymbol{T}^n A : n \geq 0\}. \end{array}$$

Let $M$ be the ground model $\langle S, I\rangle$ where $I(G)(C) = \mathbf{t}$ iff $C \in Y$, for every $C \in S$. Note that every sentence in $Y$ is nearly stably $\mathbf{t}$ in every $\tau_M$-sequence. So $Y \subseteq \mathbf{V}^{\mathbf{T}^\#}_M$. Also, $M$ is $(S - Y)$-neutral. 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 [8], there is a $\tau_M$-sequences $\mathscr{S}$ in which $B$ is neither nearly stably $\mathbf{t}$ nor nearly stably $\mathbf{f}$, so that $B \notin \mathbf{V}^{\mathbf{T}^\#}_M \cup \neg \mathbf{V}^{\mathbf{T}^\#}_M$. Incidentally, this falsifies the claim in [6] that "all sentences are nearly stable in all $\tau$-sequences for $M$." (p. 214)

If we accept Gupta and Belnap's informal case-by-case argument that $\mathbf{T}^*$ satisfies the GBD, then we must also accept that $\mathbf{T}^\#$ satisfies it. In that case the ground model in Example 5.2 must contain vicious reference, in Gupta and Belnap's informal theory-neutral sense. 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 $G$: $I(G)(A) = \mathbf{t}$ and $I(G)(B) = \mathbf{f}$, although each of $A$ and $B$ is unstable in some $\tau_M$-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 are not nearly stable. From a perspective that favours neither stability nor near stability—i.e., that favours neither $\mathbf{T}^*$ nor $\mathbf{T}^\#$—we maintain that the question whether the ground model in Example 5.2 has vicious reference, understood informally, 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 GBD 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 a determinate answer: Example 5.2 is of a non-Thomason model which is a borderline case of non-vicious reference.

The best we can show for a theory $\mathbf{T}$, using the informal notion of non-vicious reference, is that $\mathbf{T}$ satisfies what we will call the *weak* Gupta–Belnap desideratum: in clear intuitive cases of non-vicious reference, $\mathbf{T}$ dictates that truth behaves like a classical concept. But if we want to ask in general whether $\mathbf{T}$ 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 GBD—is an imprecise advantage. When we make the

alleged advantage precise through the MGBD, Theorem 5.1 states that a number of the fixed-point theories share the advantage. Furthermore, Gupta and Belnap's revision theory $\mathbf{T}^{\#}$ does not share the precise version of the advantage. We believe that Gupta and Belnap should reconsider the place of their no-vicious-reference-implies-truth-behaves-classically intuition, since not all revision theories satisfy the precise version of this intuition and a number of the revision theories' fixed-point rivals do.

## 6 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. And the fixed-point theories $\mathbf{T}^{lfp,\sigma 2}$, $\mathbf{T}^{gifp,\mu}$, $\mathbf{T}^{gifp,\kappa}$, $\mathbf{T}^{gifp,\sigma}$, $\mathbf{T}^{gifp,\sigma 1}$ and $\mathbf{T}^{gifp,\sigma 2}$ have this advantage. Although our desideratum is slightly different from 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 when it comes 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 now 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 modified 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 Section 5, 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 $\vDash$ between premise-sets and conclusions, and a proof-theoretic relation $\vdash$ (in fact, a number of proof-theoretic relations, depending on our choice of comprehension axiom). Certain desiderata involving the informal notion of logical consequence might be

satisfied by $\models$ 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 $\models$ 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 $\mathbf{T}^{gifp,\kappa}$ 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 present 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 the modified desideratum are otherwise less appealing than those that do not. All of the greatest-intrinsic-fixed-point theories satisfy our desideratum. But, although the greatest intrinsic fixed points has seemed a natural candidate for special attention, nowhere in the literature do we see arguments that any greatest intrinsic fixed point delivers the correct interpretation of truth. The only least-fixed-point theory, among those we have considered, that satisfies our desideratum is $\mathbf{T}^{lfp,\sigma 2}$. $\mathbf{T}^{lfp,\sigma 2}$ has had little explicit attention in the literature, and few advocates. It is, however, worth noting that a number of constructions and remarks in McGee [11] depend on $\sigma 2$ and favour $\sigma 2$ over both $\sigma 1$ and $\sigma$.

*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 $\tau_M$, 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" ([6], 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 in $M$.

An advocate of $\mathbf{T}^{lfp,\sigma 2}$ 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 $\tau_M$-sequences represents the behaviour of truth in $M$ will you want to privilege the class of ground models $M$ such that all $\tau_M$-sequences

culminate in the same fixed point. If, on the other hand, the behaviour of truth in $M$ is represented by $lfp(\sigma 2_M)$, then whether or not all $\tau_M$-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,[13] a sentence of the form $\boldsymbol{T}b$ 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. The name $b$ might refer to the sentence $\boldsymbol{T}c$, so that $\boldsymbol{T}b$ ultimately inherits its semantic content from whatever sentence is referred to by $c$. And so on.

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 [3] suggests that Kripke's [9] 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 $M = \langle D, I\rangle$, and let $\rho$ be some evaluation scheme. Let $\mathscr{S}$ be the $\rho_M$-sequence that builds $lfp(\rho_M)$ up from the null hypothesis—i.e., $\mathscr{S}_0(d) = \mathbf{n}$ for every $d \in D$. 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 $\mathscr{S}_\alpha(A) = \mathbf{t}$ or **f** and $I(a) = A$, then $\mathscr{S}_{\alpha+1}(\boldsymbol{T}a) = \mathscr{S}_\alpha(A)$. 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 $\alpha$, $A = I(a)$ is false and $B = I(b)$ 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

[13] See [1, 3, 4].

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 $\mathbf{T}^{lfp,\rho}$ seems like a good theory of truth, where the evaluation scheme $\rho$ is $\mu$ or $\kappa$.

Suppose that the ground language $\mathcal{L}$ contains no nonquote names, no function symbols and no nonlogical relation symbols. And let $M$ be any ground model for $\mathcal{L}$. Then $M$ displays no vicious reference, in our sense defined above, 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 $\mu$ or $\kappa$, 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 $\mu$ or $\kappa$, 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 $\mu$ or $\kappa$, either $\boldsymbol{T}$'$A$' or $\neg \boldsymbol{T}$'$A$' 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 apparent absence of vicious reference, $A$ seems ungrounded in our intuitive sense. We could come 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.

The story about grounding might trump any intuitions that blame truth's nonclassical behaviour on vicious reference. Indeed, we could go further and insist that there actually is vicious reference in this simple ground model after all, since the quote name '$\forall x(\boldsymbol{T}x \vee \neg \boldsymbol{T}x)$' viciously refers to the ungrounded sentence $\forall x(\boldsymbol{T}x \vee \neg \boldsymbol{T}x)$. 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 $\mathcal{L}$ 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.

## Appendix A: No Vicious Reference, Again

We consider another notion of non-vicious reference, which might be more intuitive than the notion defined in Section 4. We begin 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.

As for relations, we begin with classical one-place predicates. Each name is closely related to a one-place predicate, as "Pegasus" is to "pegasizes". Suppose that $M = \langle D, I\rangle$ is a ground model, where the extension of $G$ is $\{I(b)\}$ and the extension of $H$ is $\{I(c)\}$. 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 by 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 $\{I(b), I(c)\}$, refers non-viciously.

Generalizing, for classical one-place 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 relation symbols, we will say that an ordered $n$-tuple $\langle d_1, \ldots, d_n\rangle \in D^n$ is unproblematic iff each of $d_1, \ldots, d_n$ is. Then we can say that the $n$-place relation symbol $R$ 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 $n$-place function symbols, we combine the strategy for nonquote names and for relation symbols, in Definition A.2, below. We will assume that $\mathcal{L}$ and $\mathcal{L}^+$ are a corresponding ground language and truth language, and that $\mathbf{T}$ is one of our thirteen theories of truth.

**Definition A.1** Suppose that $M = \langle D, I\rangle$ is a classical ground model for $\mathcal{L}$.

1. $\mathbf{T}$ *dictates that the nonquote name* $b$ *refers non-viciously in* $M$ iff $I(b) \in (D - S) \cup \mathbf{V}^{\mathbf{T}}_M \cup \neg\mathbf{V}^{\mathbf{T}}_M$.
2. $\mathbf{T}$ *dictates that the n-place relation symbol* $R$ *refers non-viciously in* $M$ iff either
   a. for every $\langle d_1, \ldots, d_n\rangle$ in the extension of $R$, each $d_i \in (D - S) \cup \mathbf{V}^{\mathbf{T}}_M \cup \neg\mathbf{V}^{\mathbf{T}}_M$; or
   b. for every $\langle d_1, \ldots, d_n\rangle$ in the antiextension of $R$, each $d_i \in (D - S) \cup \mathbf{V}^{\mathbf{T}}_M \cup \neg\mathbf{V}^{\mathbf{T}}_M$.
3. $\mathbf{T}$ *dictates that the n-place function symbol* $f$ *refers non-viciously in* $M$ iff both
   a. for every $d_1, \ldots, d_n \in D$, $I(f)(d_1, \ldots, d_n) \in (D - S) \cup \mathbf{V}^{\mathbf{T}}_M \cup \neg\mathbf{V}^{\mathbf{T}}_M$; and
   b. for each $d \in D$, either
      i. for every $n$-tuple $\langle d_1, \ldots, d_n\rangle$ such that $I(f)(d_1, \ldots, d_n) = d$, each $d_i \in (D - S) \cup \mathbf{V}^{\mathbf{T}}_M \cup \neg\mathbf{V}^{\mathbf{T}}_M$; or
      ii. for every $n$-tuple $\langle d_1, \ldots, d_n\rangle$ such that $I(f)(d_1, \ldots, d_n) \neq d$, each $d_i \in (D - S) \cup \mathbf{V}^{\mathbf{T}}_M \cup \neg\mathbf{V}^{\mathbf{T}}_M$.

**Definition A.2** **T** *dictates that there is no vicious references in the ground model* $M$ iff **T** dictates that every nonquote name, every nonlogical relation symbol and every function symbol refers non-viciously in $M$.

Definition A.2 suggests a natural variant, say the MGBD′, of the modified Gupta–Belnap desideratum (see page 13): the MGBD′ is just like the MGBD, except that the new notion of *no vicious reference* is used. And we can look into which theories satisfy the MGBD′ (see below).

The remarks in the remainder of this appendix are due to José Martínez. First, it is easy to show that Definition A.2 implies Definition 4.17. Second, for any theory **T** that satisfies the MGBD, the two definitions are equivalent. For suppose that **T** is such a theory, and that **T** dictates that there is no vicious references in the ground model $M$ in the sense of Definition 4.17. Then **T** dictates that truth behaves like a classical concept in $M$. So $(D - S) \cup \mathbf{V}^{\mathbf{T}}_M \cup \neg\mathbf{V}^{\mathbf{T}}_M = D$. So, trivially, **T** dictates that there is no vicious references in the ground model $M$ in the sense of Definition A.2. Third, for $\mathbf{T} = \mathbf{T}^{lfp,\mu}$, $\mathbf{T}^{lfp,\kappa}$, $\mathbf{T}^{lfp,\sigma}$, $\mathbf{T}^{lfp,\sigma 1}$ or $\mathbf{T}^{\#}$, Definition 4.17 does not imply Definition A.2. Consider the following two examples.

*Example A.3* $\mathbf{T} = \mathbf{T}^{lfp,\mu}$ or $\mathbf{T}^{lfp,\kappa}$. Consider a ground language $\mathcal{L}$ with a two-place relation symbol $R$, and no other nonlogical vocabulary besides quote names. Let $A = \forall x \forall y Rxy$. Let $M$ be the ground model $\langle S, I\rangle$ where $I(R)(B, C) = \mathbf{t}$ iff $B = A$, for every $B, C \in S$. It is easy to prove that $M$ is $(S - Y)$-neutral, where $Y = \mathbf{V}^{\mathbf{T}}_M \cup \neg\mathbf{V}^{\mathbf{T}}_M$. So **T** dictates that there is no vicious references in the ground model $M$ in the sense of Definition 4.17. On the other hand, let $B = \forall x(\boldsymbol{T}x \vee \neg\boldsymbol{T}x)$ (any ungrounded sentence will do). Note: $I(R)(A, B) = \mathbf{t}$ and $I(R)(B, A) = \mathbf{f}$. So **T** dictates there *is* vicious references in the ground model $M$ in the sense of Definition A.2.

*Example A.4* $\mathbf{T} = \mathbf{T}^{\#}$. We modify Example 5.2 (Example 5.7 in [8]). Consider a ground language $\mathcal{L}$ with a one-place predicate $G$, a two-place relation symbol $R$, and no other nonlogical vocabulary besides quote names. Let

$$\begin{aligned} A &= \exists x(Gx \,\&\, \neg\boldsymbol{T}x) \\ B &= \exists x \exists y(Gx \,\&\, Gy \,\&\, \neg\boldsymbol{T}x \,\&\, \neg\boldsymbol{T}y \,\&\, x \neq y), \text{ and} \\ Y &= \{\boldsymbol{T}^n A : n \geq 0\}. \end{aligned}$$

Let $M$ be the ground model $\langle S, I\rangle$ where $I(G)(C) = \mathbf{t}$ iff $C \in Y$, for every $C \in S$ and where $R$ is interpreted as in Example A.3. As in Example 5.2, $Y \subseteq \mathbf{V}^{\mathbf{T}^{\#}}_M$, so that $\mathbf{T}^{\#}$ dictates that there is no vicious references in the ground model $M$ in the sense of Definition 4.17. On the other hand, as in Example 5.2, $B \notin \mathbf{V}^{\mathbf{T}^{\#}}_M \cup \neg\mathbf{V}^{\mathbf{T}^{\#}}_M$. Also $I(R)(A, B) = \mathbf{t}$ and $I(R)(B, A) = \mathbf{f}$. So **T** dictates there *is* vicious references in the ground model $M$ in the sense of Definition A.2.

The remaining cases, $\mathbf{T} = \mathbf{T}^{lfp,\sigma}$ and $\mathbf{T} = \mathbf{T}^{lfp,\sigma 1}$, can be dealt with by similarly modifying the Examples 5.10 and 5.11 in [8].

Finally, if **T** is one of our thirteen theories of truth, then **T** satisfies the MGBD iff **T** satisfies the MGBD′. First, for any theory **T** that satisfies the MGBD, we have already noted that the two definitions of *no vicious reference* are equivalent. So **T** satisfies the MGBD′. Second, there are five theories that do not satisfy the MGBD: $\mathbf{T}^{lfp,\mu}$, $\mathbf{T}^{lfp,\kappa}$, $\mathbf{T}^{lfp,\sigma}$, $\mathbf{T}^{lfp,\sigma 1}$ and $\mathbf{T}^{\#}$. Neither $\mathbf{T}^{lfp,\mu}$ nor $\mathbf{T}^{lfp,\kappa}$ satisfies the MGBD′, since $lfp(\mu_M)$ and $lfp(\kappa_M)$ are never classical. As for $\mathbf{T}^{\#}$ [$\mathbf{T}^{lfp,\sigma}$, $\mathbf{T}^{lfp,\sigma 1}$]: the ground model in Examples 5.7 [5.10, 5.11] in [8] is one in which $\mathbf{T}^{\#}$ [$\mathbf{T}^{lfp,\sigma}$, $\mathbf{T}^{lfp,\sigma 1}$] dictates that there is no vicious reference in the sense of Definition A.2, yet in which $\mathbf{T}^{\#}$ [$\mathbf{T}^{lfp,\sigma}$, $\mathbf{T}^{lfp,\sigma 1}$] does not dictate that truth behaves like a classical concept. So $\mathbf{T}^{\#}$, $\mathbf{T}^{lfp,\sigma}$ and $\mathbf{T}^{lfp,\sigma 1}$ do not satisfy the MGBD′.

## Appendix B: Nonsupervenience Interpretations of the Fixed-point Semantics

As noted in the first paragraph of Section 2, our thirteen theories of truth rely on the *supervenience of semantics*: the intuition that the interpretation of ***T*** should be determined by the interpretation of the nonsemantic names, function symbols and relation symbols, as represented by a ground model. M. Kremer [7] argues both that that Kripke [9] does not endorse this proposal, and that this proposal misinterprets the fixed-point semantics: the fixed-point semantics formalizes what M. Kremer calls the *fixed-point conception* of truth, according to which, as Kripke [9] 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." Note that, if we fix the evaluation scheme, the fixed-point conception favours no particular fixed point.

In Section 2 we defined what it is for a sentence of $\mathcal{L}^+$ to be valid in a *ground model* according to this or that theory. If we reject supervenience, the primary notion should not be validity 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 B.1** Suppose that $M$ is a ground model and that $h$ is a fixed point of $\mu_M, \kappa_M, \sigma_M, \sigma 1_M, \sigma 2_M$ or any other monotone operator on hypotheses. The sentence $A$ of $\mathcal{L}^+$ is *valid in the expanded model* $M + h$ iff $h(A) = \mathbf{t}$.

This notion is not relativized to a theory. But whether the notion of validity in $M + h$ is well-defined for a given hypothesis $h$ 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.

On any nonsupervenience fixed-point approach, the issue is not whether truth behaves like a classical concept in a ground model $M$, but rather in an expanded model $M + h$.

**Definition B.2** *Truth behaves like a classical concept in the expanded model* $M + h$ iff $h$ is a classical fixed point of $\tau_M$.

The analogue of the definition of *no vicious reference* (Definition 4.17) on the nonsupervenience fixed point approaches is as follows.

**Definition B.3** Suppose that $M$ is a ground model and that $h$ is a fixed point of $\mu_M, \kappa_M, \sigma_M, \sigma 1_M, \sigma 2_M$, or any other monotone operator on hypotheses. *There is no vicious reference in the expanded model* $M + h$ iff $M$ is $(S - Y)$-neutral where $Y \subseteq \{A : h(A) = \mathbf{t} \text{ or } h(A) = \mathbf{f}\}$.

From the nonsupervenience fixed-point perspective, we get a desideratum not on theories of truth, but rather on schemes of evaluation $\rho = \mu, \kappa, \sigma, \sigma 1$ or $\sigma 2$, which correspond to distinct nonsupervenience fixed-point approaches.

*Desideratum on a scheme* $\boldsymbol{\rho}$ For every ground model $M$ and every fixed point $h$ of $\rho_M$, if there is no vicious reference in the expanded model $M + h$ then truth behaves like a classical concept in the expanded model $M + h$.

**Theorem B.4**

(1) $\sigma 2$ *satisfies this desideratum, but*
(2) $\mu, \kappa, \sigma$ *and* $\sigma 1$ *do not.*

Theorem B.4, (2), follows from Theorem 4.21 in [8]. The proof of Theorem B.4, (1), is similar to the proof in [8] of Theorem 4.21 (2) for $\mathbf{T}^{lfp,\sigma 2}$.

**Acknowledgements** Many thanks to Anil Gupta, who responded to earlier drafts of this paper with very helpful criticisms and comments. Thanks to Michael Kremer for helpful conversations concerning both formal and methodological issues. Thanks to an anonymous referee for helpful comments on an earlier draft. And finally thanks to José Martínez for his results reported in Appendix A, and for other helpful comments. This research was generously supported by a grant from the Social Sciences and Humanities Research Council of Canada.

## References

1. Brandom, R. (1994). *Making it explicit: Reasoning, representing and discursive commitment*. Harvard University Press.
2. Cook, R. (2006). There are non-circular paradoxes (but Yablo's isn't one of them!). *The Monist, 89*, 118–149.
3. Grover, D. (1977). Inheritors and paradox. *Journal of Philosophy, 74*, 590–604.
4. Grover, D., Camp, J., & Belnap, N. (1975). A prosentential theory of truth. *Philosophical Studies, 27*, 73–125.
5. Gupta, A. (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, pp. 175–236 (page references are to this reprinting).
6. Gupta, A., & Belnap, N. (1993). *The revision theory of truth*. MIT Press.
7. Kremer, M. (1988). Kripke and the logic of truth. *Journal of Philosophical Logic, 17*, 225–278.

8. Kremer, P. (2009). Comparing fixed-point and revision theories of truth. *Journal of Philosophical Logic, 38*, 363–403.
9. Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy, 72*(19), 690–716.
10. Martin, R. L., & Woodruff, P. W. (1975). On representing 'True-in-L' in L. *Philosophia, 5*, 217–221.
11. McGee, V. (1991). *Truth, vagueness and paradox: An essay on the logic of truth*. Indianapolis: Hackett.
12. Yablo, S. (1993). Paradox without self-reference. *Analysis, 53*, 251–252.