An Objection to Hartry Field's Theory of Truth[1]

January 9, 2009

1 Methodology for Truth Theories

Every theorist of truth is faced with a methodological dilemma.

On the one hand, the concept of truth is commonly deployed in ordinary language use. Not only is the concept of truth commonly used, but its correct application is rarely if ever a subject of disagreement in non-philosophical circles, nor does its use appear to be the source of paradox or contradiction. (Disputes constantly arise about the alleged truth of this or that statement, but in such disputes the nature of truth is not usually at issue.) These ubiquitous and unproblematic uses of truth must serve as the "data" upon which truth theories are based. In order to be viable, a theory of truth must strive to be consistent with pre-theoretic usage of that concept. Tarski makes the same point in his informal characterization of his material adequacy condition on a definition of truth: "The desired definition [of *truth*] does not aim to specify the meaning of a familiar word used to denote a novel notion; on the contrary, it aims to catch hold of the actual meaning of an old notion" (Tarski 1944, 341).[2] Moreover, a successful theory of truth will venture to *explain* why we reason with and use that concept the way that we do. Such an explanation will take the form of general principles governing the correct usage of the truth predicate, which, if we are fortunate, will not only entail the validity of well-entrenched modes of reasoning with truth, but will also make sense of these ordinary inferences by reflecting their intuitive structure in formalism.

On the other hand, the very reason that philosophers have been pursuing an adequate

[1]This paper was inspired by ——'s lectures to his Fall 2008 seminar *Truth and Paradox* at ——, especially his presentation of November 20, 2008 in which he criticized Hartry Field's theory of truth. Thanks to —— for commenting on earlier drafts of this paper.

[2]The fact that Tarski finds the question of the definition of "truth" in natural language to be too vague to be worth investigating does not interfere with this point (Tarski 1944, 347). Tarski investigates the notion of truth in formalized languages, but he considers this notion to be the same as—or at least a precisification of—the notion of truth in natural language.

theory of truth for millennia is that simple theories that seem to entail all of the data of ordinary truth-reasoning (with a classical logic) run afoul of the liar and similar paradoxes. Such theories are of little use since, by *ex falso quodlibet*, they also entail the contradictions of all of the data of ordinary truth-reasoning. The liar paradox is a serious philosophical problem precisely because it shows that our ordinary modes of reasoning with the concept of truth—a concept we normally use with comfort—are logically illicit when extrapolated beyond ordinary cases. The truth theorist must provide a theory that avoids the logical pitfalls of the ordinary notion of truth.

There are, then, two domains a theory of truth must govern: the domain of ordinary, benign uses of the concept of truth, and the domain of pathological uses of that concept—uses that appear, via that rules that serve us well in the unproblematic domain, to lead to contradiction. To meet the conflicting demands of the two domains, the truth theorist's methodology must at once be descriptive and normative. A theory of truth must be descriptive of the pre-theoretic truth reasoning that serves us well in the first domain, but it must also make declarations about how we ought to alter our preconceptions about truth in order to avoid logical disaster in the second domain. It must conform as nearly as possible to the data of everyday reasoning with truth, while simultaneously explaining how paradoxical uses of truth do not undermine the rules of ordinary reasoning.

Based on this view of the proper methodology for constructing a theory of truth, I will argue that Hartry Field makes a methodological error in motivating his theory of truth.[3] Field argues that the chief advantage of his theory is that it is consistent with the so-called Naive Theory of Truth, whereas theories that maintain classical logic violate this principle. I argue that this feature of Field's theory is no advantage at all. After examining Field's construction and its motivations in §§2–3, I will object to Field's theory of truth as follows: In §4, I argue that it is poor methodology to stipulate the principles of the Naive Theory of

[3]The theory of Field's that I will be discussing is presented in Field 2008. An earlier version of the construction appears in Field 2003, and Field 2002 presents a significantly different but similarly motivated theory.

Truth—or any other principle that governs pathological as well as non-pathological uses of truth—as desiderata for a theory of truth. In §5, I argue that the Naive Theory of Truth, when situated in the context of Field's theory of truth, fails even to capture the features of ordinary truth-reasoning that supposedly inspired it. Therefore, the apparent adherence of Field's theory to the Naive Theory of Truth does not count in its favor. Since Field views this adherence as the principle advantage of his theory over others, my argument constitutes a direct challenge to Field's project.

2 Field's Construction

The theory of truth advocated in Field 2008 consists of a strong-Kleene-valued Kripkean minimal fixed-point language that has been supplemented with a non-truth-functional conditional, symbolized by "$\rightarrow$".[4] Let $\mathcal{L}$ be a first-order language interpreted by ground model $M = \langle D, I\rangle$, and let $\mathcal{L}$ be rich enough to express its own syntax, so that D contains names for the sentences of $\mathcal{L}$. Furthermore, let $\mathcal{L}^+$ be the result of adding an uninterpreted predicate "Tr(x)" and an uninterpreted connective "$\rightarrow$" to $\mathcal{L}$. Let $Z \subseteq D$ be the set of codes of sentences of $\mathcal{L}^+$ that have "$\rightarrow$" as their main connective. Using a strong Kleene valuation scheme with values 0, $\frac{1}{2}$, and 1, we proceed with the usual Kripkean fixed-point construction on $\mathcal{L}^+$,[5] with one alteration: the minimal fixed point is constructed on the basis not only of the atomic sentences of $\mathcal{L}^+$, but also on the members of Z. Given some initial valuation $\boldsymbol{v}$ of the formulas of $\mathcal{L}^+$, this construction gives us the minimal fixed point over $\boldsymbol{v}$ relative to the ground model M, written $\boldsymbol{v}^M$.

Field draws on resources from The Revision Theory of Truth to construct an interpretation of his new conditional.[6] He constructs an ordinal-length revision sequence (of

[4]For Field's own presentation of this theory, see Field 2008, especially §16.2.
[5]Kripke's construction of fixed-point languages can be found in his seminal paper, Kripke 1975.
[6]The Revision Theory of Truth is presented in Gupta & Belnap 1993.

hypothetical interpretations for the formulas of $\mathcal{L}^+$)

$$\mathcal{S}_M = \langle \boldsymbol{v}_0^M, \boldsymbol{v}_1^M, \ldots, \boldsymbol{v}_\alpha^M, \ldots \rangle$$

as follows. Let initial hypothesis $\boldsymbol{v}_0^M$ assign $\frac{1}{2}$ to every member of Z.[7] For successor ordinals,

$$\boldsymbol{v}_{\alpha+1}^M(B \to C) = \begin{cases} 1 & \text{if } \boldsymbol{v}_\alpha^M(B) \leq \boldsymbol{v}_\alpha^M(C) \\ 0 & \text{if } \boldsymbol{v}_\alpha^M(B) > \boldsymbol{v}_\alpha^M(C) \end{cases}$$

so that the following table lists the value of a conditional at a successor stage based on the values of its antecedent and consequent at the previous stage:

		$\boldsymbol{v}_\alpha^M(C)$		
$\boldsymbol{v}_{\alpha+1}^M(B \to C)$		0	$\frac{1}{2}$	1
	0	1	1	1
$\boldsymbol{v}_\alpha^M(B)$	$\frac{1}{2}$	0	1	1
	1	0	0	1

In order to explain how Field evaluates his conditional at limit stages, it will be helpful to introduce a notion from Revision Theory, namely that of stability. The intuitive idea is that a sentence is stably 0 (or 1 or $\frac{1}{2}$) on a given revision sequence if at some point in the sequence its value becomes 0 (or 1 or $\frac{1}{2}$) and then never waivers thereafter. Formally, a member of the domain $d \in D$ is *stably* n (for $n \in \{0, \frac{1}{2}, 1\}$) on revision sequence $\mathcal{S}_M$ iff

$$(\exists \alpha < \text{length}(\mathcal{S}_M))(\forall \beta)(\alpha \leq \beta < \text{length}(\mathcal{S}_M) \supset \boldsymbol{v}_\beta^M(d) = n).$$

A member of the domain is unstable on $\mathcal{S}_M$ iff it is not stably n for any $n \in \{0, \frac{1}{2}, 1\}$ (Gupta & Belnap 1993, 167). Here, then, is how Field evaluates his conditional at limit stages: For every limit ordinal λ and for all $n \in \{0, \frac{1}{2}, 1\}$, sentences that are stably n on the revision

[7] As Field notes, alternative constructions starting with a less neutral initial hypotheses (or using a non-minimal fixed-point language) are possible (Field 2008, 250).

sequence up to (but not including) $\boldsymbol{v}_\lambda^M$ are assigned value n by $\boldsymbol{v}_\lambda^M$, and sentences that are not stable on the preceding sequence are assigned $\frac{1}{2}$ at $\boldsymbol{v}_\lambda^M$.[8]

Given a sentence A, Field defines the "ultimate value" of A for ground model M, which I will write $||A||_M$, as follows

$$||A||_M = \begin{cases} 1 & \text{if } \exists\beta\forall\gamma > \beta(\boldsymbol{v}_\gamma^M(A) = 1) \\ 0 & \text{if } \exists\beta\forall\gamma > \beta(\boldsymbol{v}_\gamma^M(A) = 0) \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

Assigning every formula its ultimate value yields a three-valued semantics for the language $\mathcal{L}^+$ with fully interpreted truth predicate and $\rightarrow$-conditional. It will be helpful to work through two examples of how $\rightarrow$-conditionals are evaluated.

Example One. Consider the sentence $L \rightarrow L$ where L is the liar sentence. Since this statement is a conditional, it is assigned value $\frac{1}{2}$ at stage zero—that is, $\boldsymbol{v}_0^M(L \rightarrow L) = \frac{1}{2}$. At stage one, the first successor stage, we calculate the value of the statement based on the values of its antecedent and consequent at the previous stage. Since L is a paradoxical sentence, it has value $\frac{1}{2}$ at the minimal fixed point (as well as all other fixed points). Using the lookup table above, we can conclude that $L \rightarrow L$ will have value 1 at stage one since both of its constituent sentences have value $\frac{1}{2}$ at the previous stage. In fact, since L has value $\frac{1}{2}$ at every revision stage, $\boldsymbol{v}_\alpha^M(L \rightarrow L) = 1$ for all $\alpha > 1$. Therefore, the ultimate value $||L \rightarrow L||_M$ is 1, since after stage zero this statement never has any value other than 1.

Example Two. Let C be a sentence that denotes $\mathrm{Tr}(\ulcorner C \urcorner) \rightarrow \bot$, where $\bot$ denotes The False—some sentence that always has value 0.[9] What is the ultimate value of

$$\mathrm{Tr}(\ulcorner C \urcorner) \leftrightarrow (\mathrm{Tr}(\ulcorner C \urcorner) \rightarrow \bot),$$

[8] Following Gupta and Belnap 1993 (§5C), Field acknowledges the viability of alternative constructions that make use of different rules governing the valuation of conditionals at limit stages (Field 2008, 250).

[9] This sentence is sometimes referred to as the Curry Liar, after Haskell Curry.

the Tarski biconditional for $\mathrm{Tr}(\ulcorner C \urcorner)$? At all finite stages after stage zero, $\mathrm{Tr}(\ulcorner C \urcorner)$ will have value 0 at odd stages and value 1 at even stages. Thus $\mathrm{Tr}(\ulcorner C \urcorner)$ is unstable on the finite portion of its revision sequence, and so is assigned value $\frac{1}{2}$ at stage ω. At stage $\omega + 1$, the oscillation between 0 and 1 begins anew, and so the sentence will also be assigned value $\frac{1}{2}$ at the next limit stage, $\omega \cdot 2$. This pattern of instability recurs *ad infinitum*, and so $\mathrm{Tr}(\ulcorner C \urcorner)$ exhibits the following series of values for arbitrary finite ordinal β and arbitrary limit ordinal λ:

α	0	1	2	...	$\beta \cdot 2 + 1$	$\beta \cdot 2 + 2$	...	λ	$\lambda + 1$	$\lambda + 2$	...
$\boldsymbol{v}^{\mathbf{M}}_{\alpha}(\mathrm{Tr}(\ulcorner \mathbf{C} \urcorner))$	$\frac{1}{2}$	0	1	...	0	1	...	$\frac{1}{2}$	0	1	...

By its definition, the value of $\bot$ is 0 at every stage, and so $\boldsymbol{v}^{M}_{\beta+1}(\mathrm{Tr}(\ulcorner C \urcorner) \rightarrow \bot)$ will be 1 where $\boldsymbol{v}^{M}_{\beta}(\mathrm{Tr}(\ulcorner C \urcorner)) = 0$, and will be 0 where $\boldsymbol{v}^{M}_{\beta}(\mathrm{Tr}(\ulcorner C \urcorner)) = 1$. Thus $\mathrm{Tr}(\ulcorner C \urcorner) \rightarrow \bot$ will exhibit the same sequence of values at $\mathrm{Tr}(\ulcorner C \urcorner)$. It now becomes a simple matter to evaluate the two conjuncts of the biconditional. Since $\mathrm{Tr}(\ulcorner C \urcorner) \rightarrow \bot$ and $\mathrm{Tr}(\ulcorner C \urcorner)$ take the same value at each stage, the value of each of the conjuncts will be 1 at stage one and at every stage thereafter. Thus the ultimate value of the biconditional is 1.

As we will see in the next section, Field has designed his conditional expressly to ensure that the Tarski biconditionals for *every* sentence are valid.

3 Motivation for Field's Theory of Truth

Field argues that among the chief advantages of his theory of truth is its consistency with what he calls "naive" or "classical" truth theory, which is meant to be a distillation of ordinary modes of reasoning with truth.[10] (I should note explicitly that I do not mean to

[10] In Field 2008, "Naive Theory of truth-of" refers to a theory that has features analogous to those presented below, except that (T) and (PI) are stated in terms of a formula's being true of an object as opposed to a sentence's being true (p. 12). Later in Field 2008, the Intersubstitutivity Principle resurfaces in its sentential form, quoted below (p. 210). Field 2002 uses the phrase "classical truth theory" to refer to a view of truth that includes the instances of (T), and a slightly different brand of intersubstitutivity, which is explained in the next footnote (pp. 2, and especially 5–6). Field 2003 uses "naive theory of truth" or "naive truth theory" to refer to the same two features (pp. 139, 164, 165 etc.).

commit myself to the view that Field's Naive Theory of Truth successfully serves this purpose. In particular, when I wrote in §1 that the first responsibility of a theory of truth is to respect the data of ordinary reasoning, I did *not* intend to assert that the first responsibility of a theory of truth is to be consistent with Field's Naive Theory of Truth.)

According to Field, the Naive Theory of Truth (NTT) comprises two principles. The first is the the validity of the instances of the truth biconditionals:

(T) $\mathrm{Tr}(\ulcorner A \urcorner)$ if and only if A,

where A is a sentence and $\ulcorner A \urcorner$ is a name for A that is a member of the domain. The second is the following law, called the "Intersubstitutivity Principle":

(PI) If C and D are alike except that (in some transparent context) one has a sentence "A" where the other has "$\ulcorner A \urcorner$ is true", then one can legitimately infer D from C and C from D.[11] (Field 2008, 210)

Construing the "if and only if" in (T) as a conjunction of conditionals (that is, $\rightarrow$-conditionals), Field shows that every instance of (T) is valid, meaning that it has ultimate value 1 for all ground models. Moreover, (PI) is maintained for *any* sentences C and D, even when the sentences contain instances of the truth predicate embedded in conditionals, or when they are paradoxical.

Field believes NTT to represent essential features of ordinary reasoning about truth, and for this reason he considers consistency with NTT a desideratum for any truth theory. Field argues that his theory of truth is superior to those, like the Revision Theory of Truth, which must reject Intersubstitutivity. Those who would maintain a classical logic, such as revision theorists, must reject Intersubstitutivity because they are forced to accept all sentences of the form "$A \equiv A$" (where "$\equiv$" denotes the material biconditional), and in particular to

[11] Field equivocates between (PI) and another form of intersubstitutivity, namely one that "allows us to substitute occurrences of True($\ulcorner A \urcorner$) for occurrences of A, or visa versa, in any non-intensional context, when A is a sentence" (Field 2008, 65). Presumably, Field means that substituting occurrences of True($\ulcorner A \urcorner$) for occurrences of A, or visa versa, does not change the semantic status (true, false, paradoxical, etc.) of the sentences into which the substitution is made. I will focus on the version of the Intersubstitutivity Principle laid out in (PI). It is this version that I will also refer to using "Intersubstitutivity" with a capital letter "I".

accept $L \equiv L$ where L is the liar sentence. By the definition of the liar sentence

$$L \equiv \sim \mathrm{Tr}(\ulcorner L \urcorner),$$

and so if (PI) were allowed we could infer

$$\mathrm{Tr}(\ulcorner L \urcorner) \equiv \sim \mathrm{Tr}(\ulcorner L \urcorner),$$

which is contradictory.

Though Field does not defend it in detail, I will, for the sake of argument, grant the claim that ordinary reasoning condones the inferences whose validity is guaranteed by (PI). Nevertheless, it will be instructive to examine the evidence Field *does* provide for this claim. In defense of (PI), Field argues that in ordinary usage, "Talk of truth is not *just* a means of expressing agreement and disagreement" (Field 2008, 209). This claim is meant as an attack against revision theorists (and others) who must deny (PI) but can maintain the weaker principle that, for every sentence A, the circumstances under which it is proper to assent to (dissent from) A are identical to those under which it is proper to assent to (dissent from) $\mathrm{Tr}(\ulcorner A \urcorner)$. Field argues that this weaker principle does not capture sufficiently the role of truth in ordinary reasoning. What this weaker principle misses is the Intersubstitutivity of truth in truth-functionally embedded contexts.

"In particular," Field writes,

> 'true' is used inside conditionals. And in order for it to serve its purpose, it needs to be well-behaved there: inside conditionals as in unembedded contexts, 'true' needs to serve as a device of infinite conjunction or disjunction (or more accurately, a device of quantification). (Field 2008, 209–10)

By way of explanation, Field gives an example of the following form: Consider the two statements,

(1) If everything that Rush Limbaugh said in October 2008 is true then Barack Obama is a secret Marxist.

(2) If A_1 and A_2 and ... and A_n then Barack Obama is a secret Marxist.[12]

where the A_1 and A_2 and ... and A_n are the sentences uttered by Limbaugh in October 2008. Field argues that we ordinarily approve of the inferences from (1) to (2) and from (2) to (1),[13] and so his theory of truth is preferable since it respects the validity of inferences of this kind.

4 The Naive Theory of Truth as a Desideratum

In the previous section, I explained that Field desires and claims to have constructed a theory of truth that abides by the two principles that comprise NTT—(PI) and the (T) biconditionals. In this section I argue that these principles do not have the proper form to serve as reasonable desiderata on a theory of truth. To ease exposition, I will focus on (PI) and explain afterward how the same argument applies to the (T) biconditionals.

As I have mentioned, Field's impetus for stipulating (PI) as a desideratum on his theory is his observation that the word "true" performs a certain function in ordinary language—namely that of (possibly infinite) conjunction and disjunction. He concludes that "in order for the notion of truth to serve its purposes," it is necessary for (PI) to hold (Field 2008, 210). Thus, given the view of the proper methodology for truth theories laid out in §1, Field's motivation for adopting (PI) as a desideratum is of exactly the right kind: A good theory of truth will abide by (PI), Field reasons, because (PI) encapsulates an essential feature of ordinary reasoning with the concept of truth. In adopting (PI) as a desideratum,

[12] One strange aspect of this example, which carries over from Field's version, is that it is an example of finitary conjunction, whereas the passage introducing it—which I have quoted, in part, above—concerns infinitary conjunction.

[13] —— has pointed out to me that there might be circumstances in which it would be reasonable to deny the inference from (1) to (2). For instance, suppose the A_i do not jointly entail the consequent, but the A_i *plus* the fact that Limbaugh uttered those statements does entail the consequent. Field's original example admits this same objection. I will grant Field the premise that inferences like that from (1) to (2) are intuitively valid.

Field presents himself as doing nothing more or less than paying heed to the data of ordinary truth reasoning. The problem is that, even if ordinary reasoning abides by (PI), (PI) cannot be *merely* a principle of ordinary reasoning. The statement of (PI) has an implicit universal quantifier out front, namely: "*For all sentences C and D of the relevant langauge*, if ..." Thus (PI) is a principle governing *all* sentences, not just sentences that occur in everyday usage.

If (PI) captures an important feature of our ordinary practice of reasoning with truth, then a restricted version of (PI) is useful to the truth theorist as a concise summary of a feature of the data of ordinary reasoning she hopes to explain with her theory. However, it is a mistake to use (PI) as Field does: as an alledged universal law of truth, the violation of which renders a proposed truth theory illegitimate. It is a mistake to extend (PI) beyond the unproblematic cases from which it was derived by assuming that it must also hold in the pathological cases that exercise philosophers. After all, the problem of truth is that *we do not know how to reason in such cases*, because when we try to do so we meet with contradiction. To use (PI) in Field's way is to uphold that principle as an *a priori* stipulation about the nature of truth, and to make any such substantive stipulation is a methodological error. Truth theories should strive to capture the data supplied by ordinary reasoning practices as elegantly as possible while avoiding inconsistency. If it turns out that the theory that best meets this goal is one that is universally consistent with (PI), then so be it. The point is that consistency with (PI) should be checked only *after* the theory has been constructed; it should not be laid out at the start as a desideratum.

The argument that I have just given has a general methodological moral: No substantive principle that governs *all* truth-predicate-containing sentences should be set forth as a desideratum for a theory of truth. Our ordinary-reasoning intuitions should not be assumed *ex hypothesi* to apply outside of their domain of origin. Facts about ordinary reasoning should be taken as just that, not as instances of general principles governing the proper use of the concept of truth across all cases. Therefore, the totality of instances of (T) should not

be used as a desideratum any more than (PI).

5 Naive Truth Theory and Ordinary Reasoning

I have just argued that Field errs in adopting the unrestricted principles of NTT as desiderata on his truth theory. However, supporters of Field's theory may rejoin that, even if the argument of §4 is sound, Field's theory still enjoys an advantage over the theories it is meant to supplant. Namely, Field's theory captures the unproblematic, ordinary instances of (PI) and (T) that *do* constitute part of the data upon which truth theories ought to be based. In this section I argue that Field's theory enjoys no such advantage, and I argue that this is the case even if we continue to assume that ordinary reasoning does in fact respect NTT. In particular, I argue that the versions of (PI) and the (T) biconditionals that Field's theory affirms cannot be reflective of ordinary reasoning. Therefore, not only does Field err in adopting the two principles of NTT as desiderata, but he also fails to construct a theory that satisfies these desiderata in the form in which they originally appeared desirable.

Recall that Field sets himself the task of constructing a theory of truth that contains the instances of (T) and respects (PI) universally. In striving to do so, Field interprets (T) as

(T) $\mathrm{Tr}(\ulcorner A \urcorner) \leftrightarrow A$,

where "$B \leftrightarrow C$" is defined as "$(B \rightarrow C)$ & $(C \rightarrow B)$", and "$\rightarrow$" is Field's revision-theoretic conditional. As already noted, Field goes on to show that (PI) holds in all $\rightarrow$-embedded contexts. Thus Field appears to have accomplished his stated objective.

However, this appearance does not hold up to closer examination. The reason that Field seeks a theory of truth meeting the two conditions is that he believes that NTT captures the essence of how we reason about truth in ordinary circumstances. (It is for this reason, after all, that Field calls these principles the Naive Theory of Truth.") If Field is to meet his stated goal, then, his theory of truth must respect forms of (PI) and the instances of

(T) that are recognizably related to ordinary truth-reasoning. In this respect, Field's theory fails; the reason for this is that *Field's revision-theoretic conditional has no analog in ordinary reasoning*. Consider the following pieces of evidence for this claim:

"$\rightarrow$" is proof-theoretically opaque. Field himself notes the failure of his conditional to respect certain principles that hold of the material conditional. In particular, the laws of Contraction and Permutation, and the rule of Conditional Proof (i.e. $\rightarrow$-Introduction) all fail:[14]

Contraction. $A \rightarrow (A \rightarrow B) \models A \rightarrow B$

Permutation. $A \rightarrow (B \rightarrow C) \models B \rightarrow (A \rightarrow C)$

Conditional Proof. From $A \models B$ infer $\models A \rightarrow B$

The fact that Field's conditional has a different logic from the material conditional is hardly news, and it would be difficult to argue that these principles codify essential rules of ordinary reasoning.[15] Rather, the failure of these principles is troubling because, in their absence, we are left with insufficient tools for reasoning *at all* with Field's conditional. Field does not give us a demonstrably complete set of principles for conducting proofs involving his conditional, let alone a natural deduction system for a logic containing it. The proof-theoretic opacity of Field's conditional is obviously undesirable in the context of formal deductions. (It is a mystery, for instance, how one should go about proving a theorem of the form $B \rightarrow C$.) More importantly, if formal machinery does not suffice to make reasoning with "$\rightarrow$" workable, it is difficult to imagine that natural language has the necessary resources to do so.

"$\rightarrow$" is non-recursive. The interpretation of Field's conditional has the following strange property: The value of an $\rightarrow$-conditional at stage α ($\alpha > 0$) is not a function of the values of its antecedent and consequence at stage α, but rather their values at the previous stage (or, if α is a limit ordinal, at all previous stages). One cannot evaluate $\boldsymbol{v}^M_\alpha(B \rightarrow C)$

[14] For a discussion of the failure of these laws, see Field 2008, section 17.4, or Field 2003, pp. 154–6.

[15] For instance, it is difficult to formulate *any* everyday sort of inference that has the form of Conditional Proof, not least because non-philosophical natural language lacks the expressive power to distinguish between "$\models$" and the conditional.

given only the values of $\boldsymbol{v}_\alpha^M(B)$ and $\boldsymbol{v}_\alpha^M(C)$. The situation is all the worse for sentences with nested conditionals. Moreover, this non-recursiveness is not an accidental property of Field's theory; it is essential to the validity of the (T) biconditionals. For instance, if $\rightarrow$-conditionals were evaluated "flatly" at each level—that is, evaluated only in terms of the values of the constituent sentences *at the same stage of evaluation*—the (T) biconditional for the paradoxical sentence $\mathrm{Tr}(\ulcorner C \urcorner)$, discussed in §2, would be a logical falsehood. This is easy enough to verify: $\mathrm{Tr}(\ulcorner C \urcorner) \rightarrow \bot$ would have value 0 at each stage where $\mathrm{Tr}(\ulcorner C \urcorner)$ had value 1, and *visa versa*. Thus each conjunct of the boconditional would have value 0 at every (non-initial) revision stage.

"$\rightarrow$" has no reading. Perhaps the most direct piece of evidence that Field's conditional does not have an analog in natural language is that Field is unable to give a natural-language reading for "$\rightarrow$." Field laments his inability to provide an interpretation for his conditional:

> It would be nice to have not only a reasonable and sufficiently powerful logic for 'if... then' that allows us to keep [naive] truth theory, but also some sort of *interpretation* of 'if... then' that "explains" the failure of contraction (and any other classical principles governing 'if... then' that need to be abandoned) ... I do not offer a serious proposal for how to understand 'if... then' in terms of which the failure of contraction could be justified... (Field 2002, 6).

Field 2008 does not offer such an interpretation either, though it omits explicit mention of this shortcoming.

"$\rightarrow$" is ad hoc. The reason that Field has a hard time finding an English translation of the conditional that he has constructed is that his motivations for constructing it as he does are not grounded in the right way in intuitions about ordinary language. This point is essentially a reprisal of the result of §4. Compare Field's use of revision-theoretic techniques to that of Gupta and Belnap. Gupta and Belnap offer motivations for a revision-theoretic definition of the truth predicate that are grounded in ordinary usage of the concept and in the kinds of pathological behavior it exhibits.[16] They offer a circular definition of truth

[16] See Gupta & Belnap 1993, especially Part I of Chapter 4.

because, they argue, the pathology of truth shares symptoms with the pathology of circular concepts generally. These motivations are independent of the implications of the theory they go on to construct. For instance, the Revision Theory of Truth happens to preserve the (T) biconditionals, but this result does not function as a *justification* for the theory, but rather as an interesting feature of it. In contrast, Field's explicit motivation for defining a revision-theoretic conditional is that doing so allows his theory to fit with NTT. The reason he defines his conditional the way he does is that doing so allows him to achieve antecedently determined results. Field's conditional is *ad hoc*. It was conceived without any regard for the function of the conditional in ordinary reasoning, and so it is no surprise that Field cannot find a home for it there.

I have argued that Field's conditional has no analog in ordinary reasoning. The significance of this thesis bears repeating in greater detail. In motivating the use of (PI) and the (T) biconditionals as desiderata for a theory of truth, Field argues that they are principles of ordinary truth-reasoning. Later, when Field sets out to demonstrate that his theory meets these requirements, what he shows is that the $\rightarrow$-versions of these principles hold in his theory. But Field is not able to demonstrate that "$\rightarrow$" has an analog in ordinary reasoning—and, as we have seen, there is no reason to think that it does. In any case, Field has not even attempted to demonstrate that instances of (T) should be read as conjunctions of two $\rightarrow$-conditionals, nor that Intersubstitutivity should hold in $\rightarrow$-embedded contexts. If there are ordinary-reasoning intuitions that are captured by NTT, these intuitions can have nothing to do with Field's conditional. We must conclude that Field's construction of a theory of truth that appears to fit with NTT is a false victory. The version of NTT Field motivates as a desideratum for his theory does not match the version he shows his theory fits with.[17]

This line of criticism suggests a more general problem with Field's theory of truth.

[17]— inspired the idea that construing the (T) biconditionals as $\rightarrow$-statements perverts their original senses.

Field rightly seeks a theory of truth that is based upon the data of ordinary reasoning. He criticizes Kripke's fixed-point approach to truth on the grounds that "the theory is too weak to carry out ordinary reasoning," notably because it "does not contain a decent conditional or biconditional" (Field 2008, 72). In particular, Field notes that Kleene three-valued logics reject excluded middle, and so must reject "$A \supset A$" (where "$\supset$" denotes the material conditional) as a general law. To remedy this shortcoming, Field suggests adding a new conditional to a Kripkean fixed-point language.[18] Field goes on to observe that using this conditional allows for the validity of $A \rightarrow A$ and, by (PI), all of the instances of (T). The problem with this approach is that $A \rightarrow A$ is no longer the trivial bit of common-sense reasoning it once was; it now contains a conditional which is utterly foreign to ordinary reasoning. In short: Field attempts to improve the ability of Kripke's construction to carry out ordinary reasoning, but he does so by supplementing that construction with a logical connective that has no analog in ordinary reasoning, and with which reasoning *at all* is difficult at best.

I have argued that the chief selling-point of Field's theory of truth—its fit with the Naive Theory of Truth—is no advantage at all. I argued in §4 that consistency with NTT is an illegitimate desideratum for a theory of truth because NTT applies to pathological as well as ordinary uses of the concept of truth. Furthermore, as I have argued in this section, Field has not even demonstrated that his theory is consistent with a weaker version of NTT that is restricted to ordinary, benign uses of truth. Field's theory distorts (PI) and (T) in such a way that they are no longer recognizable as intuitive principles governing the concept of truth.

[18]This is literally what Field does; as Gupta has observed, the $\rightarrow$-free part of Field's construction is a plain old Kripke fixed-point language.

Works Cited

Gupta, Anil and Nuel Belnap (1993). *The Revision Theory of Truth*, Cambridge, MA: MIT Press.

Field, Hartry (2002). "Saving the Truth Schema from Paradox," *Journal of Philosophical Logic* **31**: 1–27.

—— (2003). "A Revenge-Immunte Solution to the Semantic Paradoxes," *Journal of Philosophical Logic* **32**: 139–77.

—— (2008). *Saving Truth from Paradox*, New York: Oxford University Press.

Kripke, Saul (1975). "Outline of a Theory of Truth," *Journal of Philosophy* **72**: 690–716.

Tarski, Alfred (1944). "The Semantic Conception of Truth," *Philosophy and Phenomenological Research* **4**: 341–76.