Australasian Journal of Logic

On artifacts and truth-preservation

Shawn Standefer

Abstract

In *Saving Truth from Paradox*, Hartry Field presents and defends a theory of truth with a new conditional. In this paper, I present two criticisms of this theory, one concerning its assessments of validity and one concerning its treatment of truth-preservation claims. One way of adjusting the theory adequately responds to the truth-preservation criticism, at the cost of making the validity criticism worse. I show that in a restricted setting, Field has a way to respond to the validity criticism. I close with some general considerations on the use of revision-theoretic methods in theories of truth.

In his recent *Saving Truth from Paradox*, Hartry Field presents and defends a theory of truth that rejects the validity of the law of excluded middle. A key aspect of the theory is the introduction of a new conditional. The truth predicate of Field's theory obeys the *Intersubstitutivity Principle*, which says that the substitution of A with $T(\text{`}A\text{'})$, or the converse, in any extensional context does not alter semantic value.[1] Field argues that his theory has many virtues, such as validating all of the *T-sentences*, sentences of the form

$$T(\text{`}A\text{'}) \leftrightarrow A.$$

One particular virtue for which he argues is that his theory is more satisfactory than any other approach with respect to so-called truth-preservation claims, conditionals which say that if the premises of a rule are true, then the conclusion is true. Other theories must *deny*, in the sense of assert the negation of, a truth-preservation claim for some rule the theory endorses, but, Field claims, his approach does not. Rather, Field's approach *fails to*

[1] Here 'A' is a *quotation name* of the sentence A. I will explain quotation names in §1. For more on quotation names in the study of truth, see Gupta (1982), Belnap (1982), and Kremer (1988).

endorse some instances of truth-preservation claims, and failure to endorse is viewed as better than outright denial.

I will argue that Field's theory of truth has two problems.[2] The first is that it systematically asserts connections between paradoxical sentences (§2). The second is that Field's claim about truth-preservation is, in fact, false (§3). I will examine a modification of the theory that fixes the latter problem and point to some of the difficulties that arise from the former (§4). These difficulties lead to some specific considerations about the use of revision-theoretic methods in Field's theory of truth (§5) as well as more general considerations on models and artifacts in theories of truth (§6). My arguments employ some of the technical details of Field's view, so I will begin by presenting some background on Field's theory (§1).

1 Background

Field's theory of truth combines a fixed-point theory of truth with a new conditional, so I will explain each in turn. Fixed-point theories of truth were first put forward by Saul Kripke and, independently, by Robert Martin and Peter Woodruff.[3] Since Field's theory builds on Kripke's, I will describe Kripke's construction of the fixed-point. Kripke's construction builds up an interpretation of the truth predicate, a fixed-point, in a stage-by-stage manner. The construction starts with a specification of the truth values of all the sentences of the language without the truth predicate, and the sentences with the truth predicate are initially left with no interpretation. At each successor stage, sentences are added to the extension of the truth predicate if they were true at the previous stage.[4] At limit stages the extension of the truth predicate is the union of all previous stages. The construction is bound eventually to reach a stage after which no new sentences are added to the extension.

[2]Field (2014) presents a new theory that uses formal methods different from those of Field (2008). My focus here is on the earlier theory, and due to the technical differences, the criticisms in this paper will not translate directly, if at all, to the newer theory. The points made here are still interesting, because they point to some new problems in the earlier theory and they develop some possibilities and limitations of that approach.

[3]See Kripke (1975) and Martin and Woodruff (1975).

[4]The anti-extension of the truth predicate, the sentences of which it is not true, is built up analogously at each stage, adding the sentences that were false at the previous stage.

This is the fixed-point, which is used to interpret the truth predicate.[5]

Kripke's fixed-point construction will work with any logical scheme whose connectives have a certain monotonicity property.[6] One such scheme, the one Field prefers and the one on which I will focus in this paper, is the Strong Kleene scheme. Strong Kleene has three semantic values, with **t** as the sole designated value. The truth tables for the connectives are as follows.

$\sim$	
t	**f**
n	**n**
f	**t**

&	**t**	**n**	**f**
t	**t**	**n**	**f**
n	**n**	**n**	**f**
f	**f**	**f**	**f**

$\supset$	**t**	**n**	**f**
t	**t**	**n**	**f**
n	**t**	**n**	**n**
f	**t**	**t**	**t**

$\vee$	**t**	**n**	**f**
t	**t**	**t**	**t**
n	**t**	**n**	**n**
f	**t**	**n**	**f**

My discussion will not appeal to quantifiers, so I leave them aside.

Let us give some details of the fixed-point construction. We begin with a language $\mathscr{L}$ interpreted by a classical ground model $M(= \langle D, I\rangle)$. For simplicity, assume that $\mathscr{L}$ has names and predicates, but no variables or quantifiers. We extend the language to the language $\mathscr{L}^+$ by adding a truth predicate, T, and quotation names for all the sentences of $\mathscr{L}^+$. All the sentences of $\mathscr{L}^+$ are added to the domain of the model. Quotation names have the following interpretation.

$$I(\text{`A'}) = A$$

Hypotheses, which interpret the truth predicate, will be functions from sentences to the set of Strong Kleene semantic values, $\{\mathbf{t}, \mathbf{f}, \mathbf{n}\}$. There is a partial order, $\leq$, on the values, sometimes called the information ordering: $\mathbf{n} \leq \mathbf{t}$ and $\mathbf{n} \leq \mathbf{f}$. For hypotheses f and g, let $f \leq g$ iff for all sentences A, $f(A) \leq g(A)$. I will use the notation '$M + f$' for the model that is just like M except the that truth predicate is interpreted by f. The value assigned to $T(\text{`}A\text{'})$ by $M + f$ is $f(A)$.

The construction of the fixed-point proceeds in stages from an initial hypothesis f_0, such as the hypothesis that assigns all sentences **n**.[7]

[5]This is a sketch of a set-theoretic construction, as found in Kripke (1975). There is an algebraic construction, which builds up, in a similar manner, an interpretation that assigns semantic values to sentences using the truth predicate. The set-theoretic construction is easier to explain, but the algebraic approach, which I will use, will be more useful later. My presentation will follow that of Gupta and Belnap (1993) and Visser (2004).

[6]An operator O is monotonic iff for all semantic values $\mathbf{a}_0, \ldots, \mathbf{a}_n, \mathbf{b}_0, \ldots, \mathbf{b}_n$, if $\mathbf{a}_i \leq \mathbf{b}_i$ for each i, then $O(\mathbf{a}_0, \ldots, \mathbf{a}_n) \leq O(\mathbf{b}_0, \ldots, \mathbf{b}_n)$.

[7]Other initial hypotheses can be used, although there are some restrictions that prevent the use of arbitrary initial hypotheses.

- At stage 0, let $f_0(A) = \mathbf{n}$, for all sentences A.
- At successor stages $\alpha+1$, $f_{\alpha+1} = \kappa(f_\alpha)$, where $\kappa(f_\alpha)(A)$ is the value of A in $M + f_\alpha$.
- At limit stages λ, let $f_\lambda = \bigvee_{\eta<\lambda} f_\eta$.

The construction is monotonic, in the sense that if $\alpha < \beta$, then $f_\alpha \leq f_\beta$, so it eventually reaches a *fixed-point*, which is a stage α such that $\kappa(f_\alpha) = f_\alpha$. Supposing that γ is the first stage that is a fixed-point of κ, let $f = f_\gamma$. The fixed-point f is the *minimal fixed-point* over the ground model M, and it is the interpretation of the truth predicate for M.

We will define consequence over fixed-point models. $A_1, \ldots, A_n$ have B as a consequence, in symbols, $A_1, \ldots, A_n \models B$, iff in all ground models M, if $A_1, \ldots, A_n$ are all assigned $\mathbf{t}$ by $M + f$, so is B, where f is the least fixed-point over M. Similarly, B is valid, in symbols, $\models B$, iff for all models M, B is assigned $\mathbf{t}$ by $M + f$, where f is the last fixed-point.[8]

Kripke's theory of truth has many notable features. First, paradoxical sentences such as the liar, which says of itself that it is not true, are evaluated as $\mathbf{n}$, so, on Kripke's theory, a language can consistently contain both a self-referential truth predicate and vicious self-reference.[9] Second, its truth predicate obeys the Intersubstitutivity Principle. Third, there can be many fixed-points for a single starting language. The construction sketched above focuses on the minimal fixed-point, which is reached when the initial hypothesis assigns $\mathbf{n}$ to all sentences. Some classes of fixed-points yield notions of consequence that have elegant, complete proof systems.[10]

The Strong Kleene fixed-point theory has some notable defects. First, the Strong Kleene material conditional $\mathbf{n} \supset \mathbf{n}$ is evaluated as $\mathbf{n}$, so some T-sentences, such as those for liar sentences, cannot be true.[11] Second, the material conditional does not obey substitution of equivalents, which is to say

$$A \equiv B \not\models C(A) \equiv C(B).$$

[8] Consequence can be defined with respect to different classes of fixed-points, as investigated by Kremer (1988). In this paper, I will focus on minimal fixed-points.

[9] The value $\mathbf{n}$ is assigned to sentences in neither the extension nor the anti-extension of the truth predicate.

[10] See Kremer (1988) for details.

[11] I will use '$\supset$' for the material conditional and '$\rightarrow$' for Field's conditional.

A third defect was pointed out by Kripke (1975), namely that the liar sentence is in neither the extension of the truth predicate nor its anti-extension, but there is no way to say truly, in the object language, that the liar is not true. Saying that the liar is not true, or that it is neither true nor false, using only the resources of Strong Kleene logic with the truth predicate will result in a sentence that is evaluated as **n**. The fixed-point construction will not work if there is a predicate true of all and only the sentences that are in the complement of the extension of the truth predicate.[12] Kripke's theory cannot contain, in its object languages, the resources to truthfully say that the liar is, in some way, defective.

Field augments Kripke's Strong Kleene fixed-point theory of truth with a new conditional, which solves the problems of the basic theory. Field's new conditional is defined via a revision-theoretic construction, the details of which I will now sketch.[13]

We begin with a classical model M that interprets a ground language $\mathscr{L}$, which has neither a truth predicate nor Field's conditional, $\rightarrow$. For simplicity, we assume $\mathscr{L}$ contains predicates and names, but neither quantifiers nor variables. We expand $\mathscr{L}$ to language $\mathscr{L}^+$ by adding a truth predicate, quotation names, and the conditional, $\rightarrow$, and, as before, we add the sentences of $\mathscr{L}^+$ to the domain of M. Conditional sentences will be interpreted via a hypothesis h, which is a function from conditional sentences to the set $\{\mathbf{t}, \mathbf{f}, \mathbf{n}\}$. The models $M + h$ and $M + h + f$ will be just like M except that h and f will be used to interpret, respectively, conditionals and the truth predicate.

The revision proceeds in a two-step way. Given an interpretation of conditional sentences h, the least fixed-point for truth is constructed, treating all conditionals as atoms for the duration of the fixed-point construction.[14] We will denote the least fixed-point above $M + h$ by f_{M+h}. The values of sentences in the fixed-point are used to revise the interpretation of the arrow

[12]If the logic is weakened, then such a predicate can consistently be included in the language, as shown by Gupta and Martin (1984).

[13]Revision theories of truth were discovered, independently, by Anil Gupta and Hans Herzberger, with important contributions by Nuel Belnap. See Gupta (1982), Herzberger (1982), and Belnap (1982), respectively. For a detailed overview and development revision theories of truth and definitions, see Gupta and Belnap (1993).

[14]They are treated as atoms in the sense that their semantic values are determined by the hypothesis alone. The usual recursive evaluation stops at the conditional, regardless of the complexity of the antecedent and consequent.

according to the following rules for successor and limit stages. We will write $v_\alpha(A)$ for the semantic value A receives in $M + h_\alpha + f_{M+h_\alpha}$.

- $h_0(A \to B) = \mathbf{n}$
- $h_{\alpha+1}(A \to B) = \begin{cases} \mathbf{t} & \text{if } v_\alpha(A) \leq v_\alpha(B), \\ \mathbf{f} & \text{if } v_\alpha(A) > v_\alpha(B). \end{cases}$
- $h_\lambda(A \to B) = \begin{cases} \mathbf{t} & \text{if } \exists\alpha < \lambda \forall\beta(\alpha \leq \beta < \lambda \Rightarrow v_\beta(A) \leq v_\beta(B), \\ \mathbf{f} & \text{if } \exists\alpha < \lambda \forall\beta(\alpha \leq \beta < \lambda \Rightarrow v_\beta(A) > v_\beta(B), \\ \mathbf{n} & \text{otherwise.} \end{cases}$

The initial interpretation of all conditionals is **n**. At successor stages, a conditional is revised to **t** if the value of its antecedent at the previous stage is no greater than the value of its consequent at the previous stage, and otherwise the conditional is revised to **f**. Limit stages use the rule that conditionals that have stabilized go to their stable values while unstable ones are set to **n**.

The revision process eventually enters a loop of repeating interpretations for the conditional. There are stages in the loop that serve as particularly nice interpretations for the conditional. These are Field's *acceptable stages*, which have two features. First, they are, in the terminology of Gupta and Belnap (1993), *reflection stages*, which means that sentences that are stably **x** or unstable in the revision process leading up to the reflection stage are, respectively, stably **x** or unstable in the revision process continued throughout all the ordinals. Second, the semantic value of a sentence at an acceptable stage corresponds neatly to its stability in the revision sequence. I will return to this latter point in §2 and §5. Acceptable stages are nice, in the sense that they ensure that conditionals do not change the logical behavior of the other connectives.[15] The least fixed-points over acceptable stages are the models for Field's theory of truth. One can define consequence and validity with respect to the minimal fixed-points over acceptable stages, or equivalently in terms of stability, what Field calls *ultimate value*.[16] For concreteness, say $A_1, \ldots, A_n \models B$ iff for all classical ground models M, if $A_1, \ldots, A_n$ are assigned **t** in $M + h_\gamma + f_{M+h\gamma}$, so is B, for each acceptable stage γ. Similarly, B is valid, in symbols $\models B$, iff for all classical ground models M, for all

[15] See Field (2008, 257-258) for details.

[16] See Field (2008, 251-253) for ultimate value, and Field (2003) or Field (2008, Ch. 17) for more on validity.

acceptable stages γ over M, if B is assigned $\mathbf{t}$ in $M + h_\gamma + f_{M+h\gamma}$, for each acceptable stage γ. The notation for Field's consequence relation is the same as for Kripke's, however, no confusion should arise.[17]

Earlier, I pointed out three defects of Strong Kleene logic, two centered on the material conditional and one centered on the liar. Field's conditional remedies all three defects. Field's models validate all of the T-sentences as well as the rule form of substitution of equivalents. The problem with the liar was that the Strong Kleene theory cannot say that the liar is, in a sense, defective. Field's conditional can be used to define an operator 'D', read as "determinately," that fixes this problem:

$$DA =_{Df} A \,\&\, {\sim}(A \rightarrow {\sim}A).$$

Liars, such as Ta, where $a =$ '$\sim Ta$', turn out to be not determinately true, or $\sim DTa$. New liar-like sentences can be formed using the D operator, such as Td, where $d =$ '$\sim DTd$'. Td turns out to be not determinately determinately true. More pathological sentences can be formed and evaluated using iterations of the D operator. This point is an important one to which I will return, with examples, in the next section.

In addition to remedying the three defects of Kripke's theory highlighted earlier, Field's theory of truth retains one of the noted features of Kripke's theory: The truth predicate of Field's theory obeys the Intersubstitutivity Principle.

This is sufficient background on the formal aspects of Field's theory of truth. I will now proceed to my main criticisms of Field's theory.

2 Artifacts in the theory

An important role for a theory of truth is saying what arguments involving the truth predicate are valid. It is reasonable to criticize or reject a theory for providing incorrect verdicts on many arguments. A potential source of criticism is the presence of *artifacts* in a theory, erroneous verdicts based solely on features of formal constructions connected to the truth predicate and supporting devices.[18] The artifacts are intuitively incorrect verdicts provided by

[17] Field's is the only one defined for sentences with conditionals, and the two agree when there are no conditionals involved.

[18] My use of the term "artifact" for the phenomenon to be discussed is primarily based on the usage in Yaqūb (1993, Ch. 3). The first example of a similar usage that I have been

the theory.

Many theories have been criticized for the presence of artifacts. Michael Kremer offered one such criticism of Kripke's theory of truth based on the class of all minimal fixed-points. In that theory, truth-tellers, such as $b = `Tb'$, entail liars.[19,20] This is an, arguably, incorrect verdict that the theory makes, simply because it does not consider the non-minimal fixed-points that would provide counterexamples. This is one issue that motivates the use of the class of all fixed-points.

Although he did not put it in these terms, Stephen Yablo criticized Field's theory for some artifacts involving the exclusive use of minimal fixed-points.[21] Yablo points to certain seemingly odd evaluations of conditionals involving truth-tellers, liars, and related constructions. For example, $Tb \rightarrow {\sim}Tb$ comes out as valid. Yablo's artifacts are all connected to Field's use of minimal fixed-points.

Another feature that gives rise to artifacts is the eventual periodic repetition of interpretations in revision sequences that use the simple, constant limit rule of Field's construction. In a discussion of Gupta's and Herzberger's work on revision theory, Belnap says, "[The Grand Loop] is an artifact of the construction, due entirely to the fact that the same [limit rule] is used for each and every limit stage."[22] Field's construction enters into the periodic cycle of interpretations, a Grand Loop, for the very reason Belnap points out. Belnap continues, saying, "I think the Grand Loop is an artifact created by an *ad hoc* decision to adopt always a [constant limit rule] where no such constancy is called for."[23] The loop is an artifact of the construction that results in artifacts in my sense, incorrect verdicts on the validity of arguments.

Belnap criticized the constant limit rules of Gupta and Herzberger by pointing out strange results that one gets from constant limit rules. One such is the stability of the material equivalence between two distinct liar sentences. Yaqūb (1993) criticizes the revision theories proposed in Gupta (1982) and Belnap (1982) on a similar basis. Yaqūb argues that those revision theories contain too many incorrect verdicts about the logical relations between sentences involving the truth predicate.

---

able to find is Belnap (1982, 107).

[19]See Kremer (1986). This point was also made by Visser (2004).

[20]I will use 'b' as the name of a truth-teller and 'a' as the name of a liar.

[21]Yablo (2003)

[22]Belnap (1982, 107)

[23]Belnap (1982, 107)

Field's conditional, and his theory of truth, are defined with respect to a single revision sequence, given an interpretation of the initial language. This revision sequence uses a simple, constant limit rule: all unstable conditionals are assigned **n**. The interplay between the revision rule and the limit rule creates artifacts in the theory, and these artifacts constitute a problem for Field's theory, because, as in the other cases, they are incorrect verdicts on the status of arguments.[24] I take it that a similar theory without the artifacts would be better than the theory with them. The artifacts on which I will focus are all validity claims concerning conditionals.

To help make my case, I will need some definitions. First, define iterated Curry sentences c_n as follows. Define iterated arrow notation as follows.

- $A \rightarrow_0 B =_{Df} B$
- $A \rightarrow_{n+1} B =_{Df} A \rightarrow (A \rightarrow_n B)$

Next, let $\perp$ be any contradictory, ground language sentence.[25]

Definition 1 (Iterated Curry)**.** *For each $n \geq 2$, c_n names the following sentence.*

$$Tc_n \rightarrow_{n-1} \perp$$

The revision patterns of iterated Curry sentences are simple: c_n falls into the following pattern over successor stages.[26]

$$\underbrace{\mathbf{t} \ldots \mathbf{t}}_{n-1}\mathbf{f}$$

[24]There is a question concerning what the sense of incorrectness is. These evaluations are not incorrect according to the theory that makes them; indeed, from that point of view they are simply consequences. They are incorrect in the sense of being intuitively wrong and needing philosophical justification. The claim that truth-tellers entail liars is an example. One might, following the suggestion of Yablo (2003, 319, fn. 10), defend this claim by arguing that truth-tellers are, and must be, false because nothing makes them true, and consequently there can be no counterexamples to the disputed validity claim. The dividing line between fatal flaws, artifacts, (merely) surprising consequences, and (desirable) features of a theory is to some degree fuzzy and can be a matter for philosophical debate, as, for example, the exchange between Cook (2002, 2003) and Kremer (2002) illustrates.

[25]Any falsehood of the syntactic theory, such as 'A' $\neq$ 'A', will do as long as the syntactic theory is interpreted the same way in all models.

[26]The subscript corresponds to the period of the pattern, rather than the number of nested arrows. This makes the statement of later propositions more straightforward.

At stage zero and at limit stages, the pattern for the iterated Curry sentences substitute the first value of their patterns with **n** before falling into the patterns above at subsequent successor stages.

We can, of course, define other sequences of iterated Curry sentences by changing the false sentence in the inner-most consequent. This will not affect the revision pattern, but it will generate new sequences of iterated Curry sentences.

We can define an analogous sequence of iterated determinate liars. Let D^0A be A and let $D^{n+1}A$ be $D(D^nA)$.

Definition 2 (Iterated determinate liar)**.** *For each $n \geq 1$, d_n names the following sentence.*

$$\sim D^{\,n-1}\, Td_n$$

In particular, d_1 is simply the liar, $\sim Ta$. The determinate liars also have a simple pattern of revision. For each $n \geq 2$, the pattern of revision for d_n repeats the following.

$$\mathbf{n}\underbrace{\mathbf{t}\ldots\mathbf{t}}_{n-1}$$

Henceforth, we will assume that the language $\mathscr{L}$ contains c_n and d_n for each $n > 1$, as well as a truth-teller and a liar.

The Curry sentences each comprise a name, the truth predicate, arrows, and some false sentence. The false sentence can either be a falsity constant or a contradiction made of sentences from the base language. The iterated Curry sentences will be evaluated as **n** in all acceptable stages of the construction. The iterated determinate liars each comprise a name, the truth predicate, negation, conjunction, and the arrow, so they involve even less non-logical material. They will also be evaluated as **n** in all acceptable stages. In fact, the iterated Curry and determinate liar sentences do not change their interpretation between models. In light of this, we have, for all $k \geq 2$, the following entailments, for arbitrary A.

- $Tc_k \models A$
- $Td_k \models A$

This is somewhat expected, given that the consequence relation is defined as preservation of the semantic value **t** in certain fixed-points, and the pathological sentences indicated are guaranteed not to take that value in those fixed-points.

The artifacts of Field's theory can be seen by considering *conditionals* involving the iterated Curry and determinate liar sentences, rather than entailments in in which they feature as premises. A simple calculation shows that $Tc_2 \to Tc_4$ is valid. In fact, for any even $k > 0$, $Tc_2 \to Tc_k$ is valid.

Proposition (1). *For $n, m > 1$, if $\exists k(m \cdot k = n)$, then $Tc_m \to Tc_n$ is valid.*

A similar point holds for the iterated determinate liars.

Proposition (2). *For $n, m > 1$, if $\exists k(m \cdot k = n)$, then $Td_m \to Td_n$ is valid.*

Let us say that conditionals are the *arrow correlates* of the consequence statements that are formed by replacing the main arrow of the former with a turnstile, '$\models$'.

There are then many valid conditionals that have distinct iterated Curry sentences in their antecedents and consequents. Note, however, that not all true consequence statements are reflected by valid arrow correlates, such as the following.

(i) $Tc_4 \models Tc_2$

(ii) $\not\models Tc_4 \to Tc_2$

(iii) $\not\models \sim(Tc_4 \to Tc_2)$

While (i) is a true consequence statement, neither its arrow correlate nor the negation of its arrow correlate is valid.

Field's conditional obeys *modus ponens*, so if we have a valid conditional, then corresponding consequence statement will be true. The converse is not true. As is well known, the validity of

$$A \,\&\, (A \to B) \to B$$[27]

together with a truth predicate obeying the Intersubstitutivity Principle and a conditional obeying *modus ponens* leads to triviality.[28] In Field's theory,

[27] We will adopt the convention that conjunctions and disjunctions bind more tightly than conditionals.

[28] There is an extensive literature on triviality results connected to the conditional form of *modus ponens*, $A \,\&\, (A \to B) \to B$, and its relatives. See, for example, Meyer et al. (1979), Restall (1993), Rogerson and Restall (2004), Priest (2006), Rogerson (2007), Beall (2009), Zardini (2011), or Beall and Murzi (2013).

many true consequence statements with paradoxical premises do not have valid arrow correlates.

In addition to many valid conditionals with distinct iterated Curry and determinate liar sentences, there are also many valid conditionals containing the negations of iterated Curry sentences in their antecedents and determinate liars in their consequents.

Proposition (3). *For $n, m > 1$, if $\exists k(m \cdot k = n)$, then $\sim Tc_m \rightarrow Td_n$ is valid.*

The result of switching the antecedent and consequent of the conditionals in the previous proposition as well as their subscripts, $Td_m \rightarrow \sim Tc_n$, will not be valid.

The validity of the conditionals in propositions (1)–(3) is fairly described as an artifact of the revision process defining Field's conditional. Let us call these conditionals *artifactual conditionals*.

The validity judgments concerning the artifactual conditionals present a problem for Field's theory because they are seemingly incorrect verdicts, similar to the other artifacts listed above. The artifactual conditionals appear arbitrary, yet they affirm systematic connections between pathological sentences, which are the sentences of interest for the logical behavior of Field's conditional.[29] Without attention to the details of the models, it is mysterious why a given artifactual conditional should hold, but Field attaches little philosophical significance to the models beyond demonstrating the theory's consistency.[30]

At this point, there is a response that I should address. This response says that I am illicitly imposing a requirement of *relevance* on Field's theory when he adopts no such principle.[31] The reason that the validity of the artifactual conditionals seems to be a defect is that the parts of the artifactual conditionals are irrelevant to each other. Since Field does not endorse a principle of relevance, the response concludes that my criticisms have no bite.

While I agree that the parts of the artifactual conditionals are not relevant to each other, I am not attributing an endorsement of relevance to Field.

[29] At least, they are once one notes that Field's conditional reduces to the classical material conditional under the assumption of excluded middle for antecedent and consequent.

[30] Field says, for example, "My ultimate interest is less in the semantics than in the logic that the semantics validates." Field (2008, 232)

[31] See, for example, Anderson and Belnap (1975), Read (1988), Dunn and Restall (2002), or Mares (2004) for more on relevance.

Field's theory affirms principles that violate relevance strictures, such as

$$A \models B \to A.$$

No issue is being taken with that feature of the theory. The problem with the artifacts is that their validity seems incorrect, much like the truth-teller entailing the liar in a version of Kripke's theory. The irrelevance is, perhaps, a symptom of the issue, but not the issue itself.

In §4 I will consider another response to the issue of artifacts, but for now, let us turn to truth-preservation.

3 Truth-preservation

Field (2006, 2008) stresses the importance of truth-preservation claims for the rules under which a theory is closed.[32] In the terminology of §2, the truth-preservation claim for the rule

$$A_1, \ldots, A_n \models B$$

is the validity of its arrow correlate,

$$\models T(\text{`}A_1\text{'}) \,\&\, \ldots \,\&\, T(\text{`}A_n\text{'}) \to T(\text{`}B\text{'}) \,.$$

Validity is standardly defined in terms of necessary truth-preservation. Field thinks that the truth predicate is logical vocabulary, so that validity should be broadened to encompass the use of truth and some syntactic theory. If a theory of truth accepts a rule, in the sense of being closed under that rule, then the rule is valid by lights of the theory, even if the theory does not have a validity predicate. If a theory asserts the negation of truth-preservation for one of its rules, then, even if no outright contradiction results, philosophical tension arises between the endorsement of a rule and the negation of the claim that the rule preserves truth. I will not take issue with Field's arguments for the importance of truth-preservation for theories of truth.

Field's theory of truth validates the truth-preservation claims for the truth rules, as these are T-sentences. His theory validates instances of other

[32] The earlier of the two sources places greater importance on this, although the later source does use failure of the truth-preservation as an objection against many theories. See Field (2008, Ch. 26) for the use of failure of truth-preservation as a criticism of a theory.

rules under which his theory is closed, but there are some rules that have instances that are invalid. For example, while the theory is closed under *modus ponens*, its truth-preservation claim is invalid.

$$\not\models T(\text{`}A\text{'}) \,\&\, T(\text{`}A \to B\text{'}) \to T(\text{`}B\text{'})$$

Field claims that his theory does not entail any counterexamples or disjunctions of counterexamples to any particular instance of a rule.[33]

Field's theory accepts the *ex falso* rule.

$$A, \sim A \models B$$

Its arrow correlate, however, is not valid. In fact, there are contravalid instances. For example, let Ta be a liar sentence and substitute that for A and substitute a falsehood or contradiction of the syntactic theory for B. We have the following.[34]

$$\models \sim(Ta \,\&\, \sim Ta \to a \neq a)$$

Since Field's truth predicate obeys the Intersubstitutivity Principle, this is equivalent to the negation of the truth-preservation claim for *ex falso*.[35] The problem is that Field's theory entails the rejection of the *ex falso* rule that Field wants for his logic. By Field's own lights, this is a flaw for a theory of truth.

Field's theory cannot assert truth-preservation for *ex falso*. Using logical principles accepted by the theory, truth-preservation for *ex falso* implies all instances excluded middle. Asserting truth-preservation for *ex falso* would trivialize Field's theory, as its response to the paradoxes is to reject excluded middle except when it is posited as an additional axiom for sentences containing only safe vocabulary.

[33]Field (2006, 590-591)

[34]The antecedent is bound to take the value **n** and the consequent **f**, resulting in a false conditional at each stage. This example leads to a similar problem with a version of bivalence, when quantifiers are in the language. Assume that there is a unary predicate, $Sent(x)$, added to the language when it is expanded with quotation names and that it is interpreted so as to be true of all and only sentences in the language. Then, $\sim\forall x(Sent(x) \to Tx \vee \sim Tx)$, which is a plausible rendering of the claim that not all sentences are either true or false, will be valid. Indeed, the liar provides the falsifying instance, this time with the antecedent receiving the value **t** and the consequent **n**.

[35]Rather than the arrow correlate of *ex falso*, the definition of Field's conditional ensures that $A \,\&\, \sim A \to B \vee \sim B$ is valid. This is the arrow correlate of a weakened *ex falso* rule under which Field's theory is closed.

Field's theory of truth does not say that its rules are truth-preserving. In fact, it says that one of its rules does not preserve truth. This is a failing of Field's theory that, if left unresolved, would narrow the philosophical gap between it and other theories of truth. Additionally, it throws into relief the differences between the consequence relation of Field's logic and the logical behavior of the conditional. The rule of *ex falso* is valid for Field's theory, while its arrow correlate has contravalid instances. Since the theory cannot consistently affirm truth-preservation for *ex falso*, the best that theory can do is to refrain from affirming either the truth-preservation claim or its negation. Let us turn to an option for doing so.

4 Responses

One way to fix the truth-preservation problem is to change the truth table for the revision rule of the arrow to the following.[36]

$\rightarrow$	**t**	**n**	**f**
t	**t**	**n**	**f**
n	**t**	**t**	**n**
f	**t**	**t**	**t**

This differs from the previous table in setting

$$\mathbf{n} \rightarrow \mathbf{f} = \mathbf{t} \rightarrow \mathbf{n} = \mathbf{n}.^{37}$$

This will assign **n**, rather than **f**, to the highlighted instance of *ex falso*, rendering both the arrow correlate and its negation invalid.

This modification does not, however, fix the artifact problem. The truth-table changes the revision pattern for some of the paradoxical sentences. For example, the iterated Curry sentences will have the following pattern.

$$\mathbf{n}\underbrace{\mathbf{t}\ldots\mathbf{t}}_{n-2}\mathbf{n}$$

The determinate liars will be unchanged. The upshot is that, while the artifactual conditionals in proposition (3) will be invalid, new artifactual conditionals will be valid, in addition to those in proposition (1) and (2), which

[36] I am grateful to Hartry Field for suggesting, in conversation, that I investigate the use of this truth table in the revision rule for the conditional.

[37] The limit rule must be amended to set stably **n** conditionals to **n**.

remain valid. For example, for $m, n > 1$

$$\models Tc_m \rightarrow Td_n,$$

when m evenly divides n. One additional new validity is worth noting: the simple Curry becomes equivalent to the liar.[38]

$$\models Tc_2 \leftrightarrow Ta$$

This equivalence is surprising for two reasons. First, some proponents of non-classical logic take Curry's paradox to be importantly different from the liar paradox.[39] Second, '$\sim$' is not logically equivalent to '$\rightarrow\bot$', as it is when the arrow is replaced by the Strong Kleene material conditional.

While changing the truth table for the revision rule fixes the problem with truth-preservation, it does not help with the artifactual conditionals. Some artifactual conditionals will be invalidated, but others will become valid. There is one more response that I will consider, what I will call the *axiomatization response*.

This response follows on some comments Field makes concerning the complexity of his official consequence relation.[40] Field says the following.

> [T]he set of "logically valid inferences" will have an extremely high degree of non-computability.... It might be better to adopt the view that what is validated by a given version of the formal semantics outruns "real validity": that the genuine logical validities are some effectively generable subset of those inferences that preserve value [**t**] in the given semantics.[41]

The axiomatization response says that real validity is some axiomatizable subset of the consequence relation, $\models$, and the artifacts to which I point are

[38] Many conditionals with the liar as antecedent are valid as well. For example, for all n, $Ta \rightarrow Tc_n$ and $Ta \rightarrow Td_n$ are valid.

[39] For example, the liar does, while the Curry sentence does not, fall under Priest's Inclosure Schema. They are, at least for Priest, two importantly different paradoxes. See Priest (2003, 185-186) for discussion. Field (2008), at least initially, takes Curry's paradox to motivate rejection of contraction for the conditional, and takes the liar to motivate the rejection of excluded middle. Neither direction of the equivalence is valid using the revision rule of §1.

[40] Results concerning the complexity of the consequence relation can be found in Welch (2008) and McGee (2010).

[41] Field (2008, 277)

not consequences of that axiomatization. The artifactual conditionals are, then, not really valid. Thus, my criticism has no force.

The problem for the axiomatization response is that it risks losing the nice behavior of the determinateness operator. The models ensure that for any paradoxical sentence, there is some α for which the sentence is not α-determinately true.[42] This is rightly touted as a positive feature and major achievement of the theory.[43] The challenge for an axiomatization is to categorize the defective, or indeterminate, sentences as such.[44] For some sentences, such as liars and determinate liars, it is straightforward to identify axioms to add to ensure that they are, in some sense, not determinately true. The challenge for this response will arise in a richer setting, where there will be a complex array of iterated Curry sentences as well as pathological sentences that use quantifiers.

Although the axiomatization response would provide a way to fend off the criticism based on the presence of artifacts, it threatens to undermine a key feature of Field's view.[45] Further evaluation of this response will have to wait upon the proposed axiomatization.

There is on more option to note. In a response to Yablo, Field proposes modifying his conditional's revision rule to take into account all fixed-points at each stage.[46] This modification will not affect the status of the artifactual conditionals I highlight. Perhaps the most promising option for fixing the problem with artifacts is to change the limit policy.[47] Changing the limit policy will eliminate some of the artifactual sentences, but potentially at a

[42]This is true in the restricted setting of this paper. When richer languages are under consideration, it will need some caveats, in light of the counterexamples of Horsten et al. (2012) and Welch (2014).

[43]See Field (2008, 276), for example.

[44]We will set aside the empirical case and restrict attention, as Field (2008) does, to the syntactic theory, arithmetic, or set theory.

[45]The proponent of the axiomatization response may view the situation in a more positive light. The criticism based on artifacts provides further motivation for axiomatizing Field's theory, or some subtheory thereof. The proponent may take the models to provide consistency results for possible axiomatizations and to help her see the sorts of axioms that she may be missing, namely axioms asserting that certain sentences are not determinately true, for some number of iterations of the determinately operator. I thank Graham Leach-Krouse for this suggestion.

[46]See Field (2008, 17.5) for the details.

[47]One can also change the initial evaluation of conditional sentences, but that option is less important for present purposes, since none of the artifacts are heavily dependent on the initial evaluation.

cost. To explain that cost, I must go into some of the details of reflection stages.

5 Limits and artifacts

The models for Field's theory are the minimal fixed-points over acceptable stages, which are reflection stages. As mentioned, sentences that are unstable leading up to a reflection stage are unstable in the revision sequence carried on through all the ordinals. Field's limit stage policy ensures that the semantic values of sentences at reflection stages correspond to their revision patterns.[48] Sentences with the semantic value **t** [**f**] at such stages have stabilized to **t** [**f**] in the revision process, while those that have the value **n** are unstable in the revision process.

Changing the limit policy can break the correspondence between semantic value at a reflection stage and stability in the whole revision process. Here is an example. Suppose one adopts the policy that all unstable conditionals are set to **n** at limits, except $Tc_3 \rightarrow (Tc_3 \rightarrow \bot)$, which is set to **t**. This will invalidate artifactual conditionals whose antecedent is Tc_3. However, at reflection stages, which are always limit stages, $Tc_3 \rightarrow (Tc_3 \rightarrow \bot)$ will have the semantic value **t**, as will Tc_3, while $Tc_3 \rightarrow \bot$ will not. If validity is defined as preservation of **t** over certain reflection stages, *modus ponens* will then be invalid.

We may define validity in terms of stability in the revision sequence, Field's ultimate value. If we do so, then the previous counterexample to *modus ponens* will no longer be a counterexample, as Tc_3 is not stably **t**. The danger, however, is that there is no longer a guarantee that the truth-functional connectives will behave appropriately. This is because Field's proof showing that the truth-functional connectives behave appropriately uses his limit rule in an important way.[49]

The foregoing suggests that the limit rule needs to do different things to unstable sentences at different limit stages. I have focused on a relatively simple language, one whose only pathological sentences, apart from a truth-teller and liar, are iterated Curry and iterated determinate liar sentences.

[48] See the proof of Field's Fundamental Theorem, Field (2008, 257-258).

[49] If the original truth table for the conditional's revision rule is used, then the problem will be compounded. In that situation, using the previous limit rule results in $Tc_3 \vee \sim Tc_3$ stabilizing to **t**, even though neither disjunct stabilizes.

In this context, we can improve upon Field's limit rule. First, assign each Curry, apart from c_2, and the most complex conditional in each determinate liar sentence a distinct natural number greater than 0, say, evens for Curry sentences and odds for conditionals from determinate liar sentences. Next, we require that, at successor limits, ordinals of the form $\lambda + \omega \cdot n$ for $n \geq 1$, assign **f** or **t** to sentence number n, depending on whether it is a Curry or a liar, and **n** to all other unstable sentences. At all other limit stages, assign unstable sentences **n**.

With the above limit rule, one can show that artifactual conditionals will change from **t** to **n** and back once every ω^ω stages. Thus, they will be unstable, and so be set to **n** every $\omega^{\omega \cdot \omega}$ stages. Using these facts, one can show that sufficiently long revision sequences have reflection stages that are acceptable in Field's sense. Thus, if a disjunction has ultimate value **t**, then so must one of the disjuncts. This property is needed to ensure that the logical behavior of disjunction in Field's theory matches that of Strong Kleene disjunction.

We can, then, solve the problem with the artifactual conditionals that I have highlighted, at least in this restricted setting. The language on which we have focused lacks the resources of the languages in which Field is primarily interested. In particular, it lacks quantifiers and the richness of the syntactic theory of arithmetic. In languages with arithmetic, there will be many more pathological sentences, including ones with transfinite periods in the revision sequences and there will, consequently, be a broader class of artifactual conditionals. There is, then, a question of how to define more general limit policies that invalidate the artifactual conditionals that will arise in richer settings while, at the same time, ensuring that the logic of the truth-functional connectives is not disturbed. This leads to the general conclusions.

6 Conclusion

Field uses a combination of revision sequences and fixed-point constructions to define a class of models. One can view these constructions as using the revision sequences to generate semantic values.[50] Suppose that the period of the revision process between acceptable stages is Σ. We can view each Σ-long sequence from $\{\mathbf{t}, \mathbf{n}, \mathbf{f}\}$ as a possible semantic value, ordered point-wise. These values are partially, not linearly, ordered. When the ordering relation

[50] For more on this idea, see Field (2008, 259-262) and Priest (2010), especially §4.

holds between the values assigned to sentences A and B, then $A \rightarrow B$ receives the top value, which corresponds to stably **t**, the sole designated value.

Despite the multitude of potential values, the revision process defining the conditional constrains which values can be assigned to conditionals, so the value assigned to a paradoxical conditional, such as an iterated Curry, stands in the ordering relation to the values of many other paradoxical sentences. Given a ground model, initial evaluation, and constant limit rule, there is only one value that can be assigned to the conditionals. Indeed, since the interpretations of the parts of the artifactual conditionals do not change between models, these sentences can receive just one value, just one possible sequence in the revision process, in all models. In the simplified setting of this paper, the use of the limit rule of §2, in a sense, reduces the transfinite revision period, Σ, to a finite period for each sentence. These factors, a single interpretation and a reduction to finite periods, combine to ensure that there are many artifactual conditionals.

At this point, a comparison with Kripke's Strong Kleene theory will be helpful. In Kripke's theory, there is only one value that the liar can receive, **n**. Truth-tellers, as Tb, can receive, by contrast, any of the three values, **t**, **f**, or **n**. If one considers only minimal fixed-points, then

$$Tb \models Ta,$$

and for all truth-tellers Tb_1 and Tb_2,

$$Tb_1 \models Tb_2 \text{ and } Tb_2 \models Tb_1.$$

If one considers all fixed-points, however, then all of the above entailments fall away. The result is simpler, although correspondingly weaker, logic.

Similarly, when using a revision-theoretic construction to define a logic, one needs to consider non-constant limit rules. Failure to do so ensures that the resulting theory will make apparently wrong claims about valid arguments and sentences. In the case of the $S^{\#}$ revision theory of truth, consideration of a broad class of limit rules eliminates many of the unappealing validities found in some revision theories.[51] The overall logic, including the truth predicate, is arguably better for it.

I have presented two problems for the theory of truth of Field (2008). The truth-preservation problem can be fixed by adjusting the definition of

[51] See Gupta and Belnap (1993, 218-229) for more on the $S^{\#}$ theory of truth.

the conditional, an adjustment which does not fix the artifact problem. In the restricted setting I examine, the artifact problem can be fixed by adopting a different limit rule, in particular a non-constant rule, although some care must be exercised in choosing the rule. In a richer setting, further changes to the limit rule will be needed, although that presents a potential hurdle: some limit rules disrupt the disjunction property above, namely, that if a disjunction receives ultimate value $\mathbf{t}$, then so must one of its disjuncts. I will leave open the question of what rules one might adopt in a richer setting.

Artifacts in a theory of truth point to ways in which models fail to provide sufficient variation in the interpretation of sentences. Identifying ways to eliminate artifacts, when possible, points to potential avenues for improving the theory, as well as general limitations of the approach.

Acknowledgements

I am very grateful to Anil Gupta and James Shaw for discussing this work with me. I would also like to thank Graham Leach-Krouse and the anonymous referee of this journal for helpful feedback.

References

Anderson, A. and Belnap, N. (1975). *Entailment: The Logic of Relevance and Necessity*, volume 1. Princeton University Press.

Beall, J. (2009). *Spandrels of Truth*. Oxford University Press.

Beall, J. and Murzi, J. (2013). Two flavors of Curry's paradox. *Journal of Philosophy*, 110(3):143–165.

Belnap, N. (1982). Gupta's rule of revision theory of truth. *Journal of Philosophical Logic*, 11(1):103–116.

Cook, R. (2002). Counterintuitive consequences of the revision theory of truth. *Analysis*, 62(273):16–22.

Cook, R. (2003). Still counterintuitive: A reply to Kremer. *Analysis*, 63(279):257–261.

Dunn, J. M. and Restall, G. (2002). Relevance logic. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, pages 1–136. Kluwer.

Field, H. (2003). A revenge-immune solution to the semantic paradoxes. *Journal of Philosophical Logic*, 32:139–177.

Field, H. (2006). Truth and the unprovability of consistency. *Mind*, 115(459):567–605.

Field, H. (2008). *Saving Truth from Paradox*. Oxford.

Field, H. (2014). Naive truth and restricted quantification: Saving truth a whole lot better. *Review of Symbolic Logic*, 7(1):147–191.

Gupta, A. (1982). Truth and paradox. *Journal of Philosophical Logic*, 11(1):1–60.

Gupta, A. and Belnap, N. (1993). *The Revision Theory of Truth*. MIT Press.

Gupta, A. and Martin, R. L. (1984). A fixed point theorem for the weak Kleene valuation scheme. *Journal of Philosophical Logic*, 13(2):131–135.

Herzberger, H. G. (1982). Notes on naive semantics. *Journal of Philosophical Logic*, 11(1):61–102.

Horsten, L., Leigh, G. E., Leitgeb, H., and Welch, P. (2012). Revision revisited. *Review of Symbolic Logic*, 5(4):642–665.

Kremer, M. (1986). *Logic and Truth*. PhD thesis, University of Pittsburgh.

Kremer, M. (1988). Kripke and the logic of truth. *Journal of Philosophical Logic*, 17:225–278.

Kremer, M. (2002). Intuitive consequences of the revision theory of truth. *Analysis*, 62(4):330–336.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72:690–716.

Mares, E. D. (2004). *Relevant Logic: A Philosophical Interpretation*. Cambridge Univeristy Press.

Martin, R. L. and Woodruff, P. W. (1975). On representing 'true-in-L' in L. *Philosophia*, 5(3):213–217.

McGee, V. (2010). Field's logic of truth. *Philosophical Studies*, 147(3):421–432.

Meyer, R. K., Routley, R., and Dunn, J. M. (1979). Curry's paradox. *Analysis*, 39(3):124–128.

Priest, G. (2003). *Beyond the Limits of Thought*. Oxford University Press, 2nd edition.

Priest, G. (2006). *In Contradiction: A Study of the Transconsistent*. Oxford University Press, 2nd edition.

Priest, G. (2010). Hopes fade for saving truth. *Philosophy*, 85(1):109–140.

Read, S. (1988). *Relevant Logic: A Philosophical Examination of Inference*. Blackwell.

Restall, G. (1993). How to be *really* contraction free. *Studia Logica*, 52(3):381–389.

Rogerson, S. (2007). Natural deduction and Curry's paradox. *Journal of Philosophical Logic*, 36(2):155–179.

Rogerson, S. and Restall, G. (2004). Routes to triviality. *Journal of Philosophical Logic*, 33(4):421–436.

Visser, A. (2004). Semantics and the liar paradox. In Gabbay, D. and Guethner, F., editors, *Handbook of Philosophical Logic*, volume 11, pages 149–240. Springer, 2nd edition.

Welch, P. D. (2008). Ultimate truth vis-à-vis stable truth. *Review of Symbolic Logic*, 1(1):126–142.

Welch, P. D. (2014). Some observations on truth hierarchies. *Review of Symbolic Logic*, 7(1):1–30.

Yablo, S. (2003). New grounds for naive truth theory. In Beall, J., editor, *Liars and Heaps: New Essays on Paradox*, pages 312–330. Oxford University Press.

Yaqūb, A. M. (1993). *The Liar Speaks the Truth: A Defense of the Revision Theory of Truth*. Oxford University Press.

Zardini, E. (2011). Truth without contra(di)ction. *The Review of Symbolic Logic*, 4(04):498–535.