J Philos Logic (2017) 46:27–63
DOI 10.1007/s10992-015-9393-3

Conditionals in Theories of Truth

Anil Gupta[1] · **Shawn Standefer**[2]

Received: 12 June 2014 / Accepted: 29 December 2015 / Published online: 1 February 2016

Abstract We argue that distinct conditionals—conditionals that are governed by different logics—are needed to formalize the rules of Truth Introduction and Truth Elimination. We show that revision theory, when enriched with the new conditionals, yields an attractive theory of truth. We go on to compare this theory with one recently proposed by Hartry Field.

Keywords Truth · Paradox · Revision theory · Conditionals · Circular definitions

1 Introduction

The conditionals we will be concerned with are those used in stating the rules for truth, Truth Introduction (TI) and Truth Elimination (TE):

(TI) If A, then 'A' is true; and
(TE) If 'A' is true, then A.

We will argue that the two conditionals here are different: they do not mean the same; they are not governed by the same logic.

✉ Anil Gupta
agupta@pitt.edu

Shawn Standefer
sstandefer@unimelb.edu.au

[1] University of Pittsburgh, Pittsburgh, PA 15260, USA

[2] University of Melbourne, Parkville, VIC, 3010, Australia

In outline, our argument will be as follows. We will take it that (TI) and (TE) are the best formulations we have of the principal rules governing truth. Aristotle suggested these rules in the *Categories*; the medievals called them 'Aristotle's Rules'; and at least on this bit of logic no one has improved on Aristotle. We will compare these rules to the inferential behavior of the truth predicate. Whatever readings of the conditionals in (TI) and (TE) allow us to make the best sense of this behavior, those are the right readings for these conditionals. We will argue that the readings that make the best sense of the inferential behavior of truth assign different meanings, and different logics, to the conditionals in (TI) and (TE).

Consider some simple examples of arguments involving 'true':

(A1) Snow is white; therefore, 'snow is white' is true.

(A2) Grass is white; therefore, 'snow is white' is true.

(A3) Suppose everything Fred says is true. Then, everything Fred says is true. So: if everything Fred says is true then everything Fred says is true.

(A4) Snow is white; therefore, if everything Fred says is true then 'everything Fred says is true and snow is white' is true.

(A5) Grass is not white; therefore, if 'grass is white or everything Fred says is true' is true then everything Fred says is true.

Arguments (A1) and (A3) are plainly valid, and (A2) invalid. Furthermore, there are readings of (A4) and (A5) on which they are valid.[1] Consider next some arguments in which paradoxical self-reference is in play:

(A6) Given: l = 'l is not true'. If l is true, then l is not true. So, if l is true, then l is true and also l is not true. We conclude, l is not true.

(A7) Given: b = 'if b is true then God exists'. Suppose, b is true. So, if b is true, then God exists. Hence, God exists. We can conclude, therefore, that if b is true then God exists. Hence, b must be true. So, God exists.

Argument (A6) is one-half of the Liar argument, and it is highly perplexing. Its status is unclear, and different logicians have pronounced differently on its validity. Argument (A7) is Curry's Paradox and it is plainly invalid, though it is debatable where precisely the argument goes wrong. Consider, finally, some arguments containing other sorts of self-reference:

(A8) Given: t = 't is true'. Suppose t is true. Then, 't is true' is true. So: if t is true then 't is true' is true.

(A9) Given: t = 't is true' and l = 'l is not true'. Suppose l is not true. Suppose further that l is true. We have a contradiction, and we may conclude t is true. So: if l is not true then t is true if l is true.

(A10) Given: t = 't is true'. Suppose t is true. Then, snow is black.

[1] Here and below, we treat quotation names as falling among the logical constants; their interpretation does not shift in assessments of validity.

Arguments (A8) and (A9) do not bring into play the rules for truth, and they appear to be valid. Argument (A10), on the other hand, seems to be invalid: the transition from the Truth-Teller, 't is true', to 'snow is black' appears to be unwarranted. The rules of logic, even when supplemented with (TI) and (TE), do not enable us to derive 'snow is black' from the Truth-Teller. Simple examples like these can, of course, be multiplied indefinitely. Moreover, less simple (and thus more interesting) examples can also be given.

There is, then, a universe of arguments involving 'true'. Some of the arguments in this universe are plainly valid; some are plainly invalid; and the status of some others is uncertain and debatable. We want to better understand the logic of these arguments. We want a systematic account that separates valid arguments from invalid ones, and we want an explanation *why* the valid arguments are valid, and the invalid ones invalid. We have here a problem of logic that is fully classical in form: a problem of understanding validity.

The problem as stated is much too hard, however. To solve it, we would need to solve all the other problems of logic (e.g., those of vagueness); for the validity of an argument containing 'true' depends sometimes not only on the rules for truth but also on the behavior of the other elements in it (e.g., vague terms). Let us assume, therefore, that apart from truth and the conditionals, nothing else is problematic in the language. Let us go further and set down the following desideratum on a logic of truth:

Descriptive Adequacy: Let $\mathscr{L}$ (the "ground" language) be a classical first-order language with the usual logical resources. For definiteness, let us take identity ($=$), negation ($\neg$), conjunction (&), and the universal quantifier ($\forall$) as primitives; and let us take disjunction ($\vee$), material conditional($\supset$), material equivalence ($\equiv$), the existential quantifier ($\exists$), and the constants The True ($\top$) and The False ($\bot$) to be defined in any one of several familiar ways. Let us assume, for simplicity, that $\mathscr{L}$ has no function symbols.[2] Let $\mathscr{L}$ be expanded to $\mathscr{L}^+$ through the addition of a truth predicate T ("true in $\mathscr{L}^+$") and possibly one or more conditionals. We assume that $\mathscr{L}$ has canonical names for the sentences of $\mathscr{L}^+$ (these names may be formed using quotation marks). Then: We want an account of truth and of the conditionals that (i) separates valid arguments of $\mathscr{L}^+$ from invalid ones in a satisfactory way; (ii) provides explanations of why the valid arguments are valid, and invalid ones are invalid; and (iii) sees the logical behavior of T as issuing from (TI) and (TE).

Descriptive Adequacy highlights classical ground languages, but there is nothing special about these languages except their familiarity. The perplexing inferential

[2] The above stipulations concerning logical resources will remain in effect for all formal languages considered below. Note that the exclusion of function symbols implies no loss of expressive power, for an $(n+1)$-ary predicate can do the expressive work of an n-ary function symbol. The elimination of function symbols is merely a convenience; nothing essential hinges on it. Note also that, here and below, we often use symbols autonymously.

behavior of truth arises in other logical contexts also: three-valued, four-valued, infinite-valued, intuitionistic, and relevance. Furthermore, concepts other than truth (e.g., “exemplification”) exhibit very similar behavior. This motivates a further desideratum:

Generality Requirement: The account of truth and conditionals should be generalizable to non-classical logics (such as, many-valued, intuitionistic, and relevance) and to other concepts (such as “exemplification” and “satisfaction”).

These are the principal desiderata we aim to respect as we work out a logic of truth and conditionals.[3] We will argue that an enriched version of revision theory is well placed to meet the desiderata. Revision theory provides general schemes for making sense of circular and interdependent definitions. These schemes do not presuppose a specific logic; hence, revision theory is well placed to meet the Generality Requirement. Furthermore, the schemes yield specific theories of truth once we take truth to be a circular concept governed by Aristotle’s Rules.[4] The resulting theories of truth are attractive, but they suffer from a serious lacuna. They lack conditionals suitable for expressing (TI) and (TE). An effect of this expressive shortcoming is that these theories are unable to account for a large class of arguments. For example, they are unable to provide readings under which (A4) and (A5) are valid. Our aims in this essay are (i) to show how the requisite conditionals can be added to revision theory and (ii) to make a case that the resulting theory meets the requirement of Descriptive Adequacy. It will turn out that the conditionals needed to express (TI) and (TE) differ from one another in their logic and semantics. It will turn out also that the logical behavior of the conditionals is quite extraordinary.

We begin by showing that the conditionals in (TI) and (TE) are not ordinary conditionals. Then we will present an extended revision theory capable of expressing (TI) and (TE), and we will compare this theory with another which also deems these conditionals to be extraordinary.

2 Ordinary Conditionals

We argue that the conditionals in Aristotle’s Rules for truth are not ordinary conditionals. We begin by showing that, in $\mathscr{L}^+$, these conditionals should not be read as the classical material conditional. We will then argue that a shift to a non-classical material conditional does not provide a satisfactory reading, either. This will put us in a position to conclude that the conditionals in Aristotle’s Rules are not ordinary conditionals, in the sense made precise below.

[3] Other desiderata have been proposed for theories of truth. For a discussion of these, see Gupta [31], McGee [47], Simmons [64], Maudlin [46], Priest [55], and Shaw [63]. See Leitgeb [41] for an overview of consistent and inconsistent sets of desiderata.

[4] Revision theories of truth were discovered, independently, by Anil Gupta and Hans Herzberger, with seminal contributions by Nuel Belnap. See Gupta [28], Herzberger [36], and Belnap [8]; see also Yaqūb [79] and Chapuis [16]. The idea that truth is a circular concept, of which Aristotle’s Rules are partial definitions, was first put forward in Gupta [29].

2.1 Classical Material Conditional

If the conditionals in the rules for truth are interpreted as the classical material conditional, then the rules read thus:

$$T(\text{‘}A\text{’}) \supset A, \text{ and } A \supset T(\text{‘}A\text{’}).$$

Under this reading, the rules for truth are inconsistent if a liar sentence is expressible in the language.[5] This consequence has been embraced by some philosophers, most notably by Charles Chihara. Chihara, following a suggestion of Tarski, defends the above reading of the rules of truth and the resulting idea that truth is an inconsistent concept.[6]

Chihara's inconsistency view has the merit that it can easily meet the Generality Requirement: it is easy enough to find readings of the conditionals under which the rules for truth are inconsistent in expressively rich non-classical languages. However, as Chihara in effect recognizes, the proposal does poorly on Descriptive Adequacy. The proposal declares arguments valid that are plainly invalid. If the premisses of an argument merely imply the existence of a liar sentence then, according to the inconsistency view, the argument is bound to be valid, irrespective of the argument's conclusion. More specifically, the inconsistency view declares argument (A7) to be valid when, intuitively, it is invalid. Chihara responds to the problem by suggesting that in working with truth, as in working with any inconsistent theory, one should not blindly follow where validity leads. One should restrict oneself to using only, what Chihara calls, "reasonable inferences." Truth, he says, is inconsistent, but this fact does not preclude it from being useful.

However, without a satisfactory demarcation of inferences that are "reasonable," Chihara's proposal is no solution to our problem; it provides no illumination of the logic of truth. What advice, for instance, can Chihara offer the ordinary user of 'true'? That in working with 'true' the user should take care to appeal only to "reasonable inferences"? This is not helpful advice unless the theorist is prepared to spell out "reasonable inference"—a task that is much harder than that of providing a logic of truth. Chihara is reducing a hard problem to a virtually impossible one. Neither Chihara nor any other inconsistency theorist we know has provided a satisfactory account of "reasonable inference" or of how we should reason with the concept of truth.[7]

[5] We are setting aside hierarchical/contextual conceptions of truth because we do not think that they are descriptively adequate, partly for reasons given in Gupta [28]. See Tarski [69–71], Parsons [51], Burge [12], Barwise and Etchemendy [6], and Glanzberg [27] for different articulations of this conception. Skyrms [65], Gaifman [25, 26], and Simmons [64] develop theories that are broadly contextual but rather different from the main-stream contextual theories.

[6] See Chihara [17, 18]. Similar ideas have been defended by, among others, Eklund [21], Patterson [52], Scharp [62], and Burgess and Burgess [13].

[7] The classical material conditional reading is problematic even when paradoxical self-reference is absent—that is, even when (TI) and (TE) yield no inconsistency. See Martin [43, p. 198].

2.2 Non-Classical Ordinary Conditionals

It is an ancient idea that paradoxical sentences are neither true nor false, and lovely contemporary theories of truth have been built on it. Saul Kripke and, independently, Robert L. Martin and Peter Woodruff proved that if the truth predicate is allowed to be partial then attractive fixed-point interpretations can be found for it. More precisely, let M (the "ground" model) be the classical interpretation associated with $\mathscr{L}$. Set $M = \langle D, I\rangle$, where D is the domain and I is the interpretation function of M. Now, suppose we allow the truth predicate T of $\mathscr{L}^+$ to be assigned partial interpretations $\langle U, V\rangle$, where $U \cap V = \emptyset$ and U and V are subsets of D. Here U represents the extension, and V the anti-extension, of the predicate. Let $M + \langle U, V\rangle$ be the model just like M except that it assigns to T the interpretation $\langle U, V\rangle$. The truth values of sentences can now be calculated using one of several three-valued semantic schemes. Let us fix on the Strong Kleene scheme, κ, and define the operation κ_M as follows:[8] $\kappa_M(\langle U, V\rangle) = \langle U', V'\rangle$, where

$$U'(V') \text{ is the set of sentences true (false) under } \kappa \text{ in } M + \langle U, V\rangle.$$

Observe that if $\langle U, V\rangle$ is a fixed point of κ_M—that is, if $\kappa_M(\langle U, V\rangle) = \langle U, V\rangle$—then:

$$A \in U \text{iff } A \text{ is true in } M + \langle U, V\rangle, \text{ and}$$
$$A \in V \text{iff } A \text{ is false in } M + \langle U, V\rangle.$$

That is, if T is assigned a fixed-point interpretation $\langle U, V\rangle$, then T agrees perfectly with what turns out to be true, and what false, in the resulting language (i.e., in the language with the interpretation $M + \langle U, V\rangle$). Kripke showed that κ_M has fixed points.[9] Furthermore, he showed that κ_M has a least fixed point, and that this fixed point captures important intuitive properties of the truth predicate.

The fixed-point theories of truth made available by the work of Kripke, Martin, and Woodruff are lovely, but they are not the final word on our problem. First, while the inconsistency view declares too *many* inferences to be valid, fixed-point theories declare too *few* of them to be valid. Thus, under the Strong Kleene scheme, a simple argument such as (A3) is invalid.[10]

(A3) Suppose everything Fred says is true. Then, everything Fred says is true. So: if everything Fred says is true then everything Fred says is true.[11]

[8] For an account of the Strong Kleene scheme and for a more detailed presentation of fixed-point theories, see Gupta and Belnap [34, ch. 2].

[9] Martin and Woodruff showed the existence of fixed points for the Weak Kleene scheme; Kripke showed that fixed points exist for all monotonic schemes, including Strong Kleene and supervaluation schemes.

[10] This point is a variant of an observation in Gupta [28].

[11] If the truth theory is based on the least fixed point of the Strong Kleene scheme, then (A8) and (A9) are also ruled invalid, while the inference in (A10) from the Truth-Teller to the Liar is deemed valid. Some of these flaws can be removed by shifting to different fixed points or by shifting valuation schemes, but only at the cost of generating other flaws. As far as we can tell, no fixed-point theory fully meets the Descriptive Adequacy requirement. This point holds for the fixed-point theories made available by Priest [53], Visser [72, 73], and Woodruff [77].

Second, the rules for truth cannot be affirmed in the fixed-point languages. Under the conditionals definable in these languages, some instances of (TI) and (TE) turn out not to be true.[12] To be sure, conditionals are available within three-valued logics which, if present, would render (TI) and (TE) true. The problem is that the addition of these conditionals destroys the fixed-point property. For example, if we add to the Strong Kleene language a conditional, $\rightsquigarrow$, that has the semantics,

$$\mathbf{n} \rightsquigarrow \mathbf{n} = \mathbf{t},$$

and that agrees with the material conditional on classical values, then we can preserve (TI) and (TE), but we lose the fixed-point property.[13]

Third, and this point is related to the previous one, fixed points exist only for expressively weak languages.[14] The problem before us, however, is that of understanding the behavior of truth in expressively rich languages. The fixed-point approach cannot be extended to these languages, and hence it fails to meet the Generality Requirement.

The above considerations motivate a general constraint on the reading of the conditionals in (TI) and (TE). Let $\Rightarrow$ be an arbitrary conditional, and let the corresponding biconditional be $\Leftrightarrow$:

$$A \Leftrightarrow B =_{Df} (A \Rightarrow B) \ \& \ (B \Rightarrow A).$$

Let us say that $\Rightarrow$ is *ordinary* iff an enrichment with a negation connective, $\neg$, is possible that renders the following a logical law:

$$\neg(A \Leftrightarrow \neg A).$$

(By this definition, a wide range of conditionals count as ordinary, including intuitionistic, relevance, and the Strong Kleene conditionals.) Now, the constraint on the reading of the conditionals in (TI) and (TE) is this: these conditionals cannot be read as one and the same ordinary conditional, $\Rightarrow$.[15] For, if we do so, then expressive richness forces the existence of a Liar sentence $\neg Tl$ for which we should affirm

$$Tl \Leftrightarrow \neg Tl.$$

But, in light of the above logical law, we should also affirm

$$\neg(Tl \Leftrightarrow \neg Tl).$$

We are thus landed in an inconsistency. Hence, if we wish to embrace expressive richness and at the same time avoid inconsistency, as we have argued we should, then there is no choice but to interpret the conditionals in (TI) and (TE) as something other than ordinary conditionals.

[12] For dialetheic languages: these conditionals do not turn out to be solely true.

[13] Martínez–Fernández [45] has characterized exhaustively the three-valued truth-functional languages that admit the fixed-point property.

[14] For discussion of expressive limitations, see Priest [54, 55], Field [23, ch. 21], and Beall [7, pp. 52-57], among others.

[15] The idea that the conditionals in (TI) and (TE) might be different *ordinary* conditionals is unmotivated, and we set it aside.

3 Step Conditionals

We now introduce two new conditionals within the framework of the revision theory of definitions and truth. As we noted above, revision theory provides general schemes for making sense of circular and interdependent definitions. These schemes yield, in turn, specific theories of truth if we take truth to be a circular concept, with Aristotle's Rules as its partial definition. Let us review some basic facts about the revision theory of definitions.[16] This will put us in a position to introduce the new conditionals and to reflect on the resulting theories of truth. Readers not wanting to work through technical details will find an informal summary account of the conditionals in Section 3.4 below.

3.1 Revision Theory of Definitions

Let $\mathscr{L}^-$ be a classical first-order language with interpretation M ($= \langle D, I\rangle$) that is extended to a language $\mathscr{L}$ through the addition of a new one-place predicate G. Let G be governed by the following definition,

$$(\mathscr{D}) \quad Gx =_{Df} A(x, G),$$

where $A(x, G)$ may contain occurrences of G but may contain no free occurrences of any variable other than x.[17] In revision theory, circular predicates are understood in terms of revision rules, rules that revise, what we shall call, *hypotheses*. We need a few preliminary concepts before we can define this notion.[18]

Let F be the set of formulas of $\mathscr{L}$ that contain no names, and let V be the set of assignments v of values to the variables (relative to M). Let us say that a pair $\langle A, v\rangle$ consisting of a formula and an assignment *is similar to* a pair $\langle B, v'\rangle$ iff both A and B have the same number, n, of free occurrences of variables and there is a formula $C(z_1, \ldots, z_n)$ with precisely n distinct free variables $z_1, \ldots, z_n$ and

(i) for some variables $x_1, \ldots, x_n$, A is an alphabetic variant of $C(x_1, \ldots, x_n)$;[19]
(ii) for some variables $y_1, \ldots, y_n$, B is an alphabetic variant of $C(y_1, \ldots, y_n)$; and
(iii) for all i, $1 \leq i \leq n$, $v(x_i) = v'(y_i)$.

[16] The review is necessarily compressed. For a fuller exposition of the theory, see Gupta and Belnap [34].

[17] Revision theory provides schemes for interpreting systems of interdependent definitions of terms of all categories. Furthermore, it is straightforwardly generalizable to systems of partial definitions. For notational simplicity, we work with a system consisting solely of one definition that circularly defines a one-place predicate G.

[18] The notion of hypothesis we need is more complex than in earlier revision theories because of the forthcoming introduction of the conditionals. Our hypotheses must provide not only interpretations for the defined term, they must also provide bases for the evaluation of the new conditionals. If the hypotheses were not suitably restricted, the behavior of the conditionals would be unacceptable. We formulate a suitable restriction below. In the absence of the new conditionals, hypotheses may be identified with subsets of the domain D.

[19] That is, A and $C(x_1, \ldots, x_n)$ are exactly like except perhaps for a difference in bound variables. We assume that substitution creates no clash of variables. So, $x_1, \ldots, x_n$ must be all and only free variables of A

Then, a *hypothesis* h is a subset of $F \times V$ that satisfies the following condition:

$$\text{for all similar pairs } \langle A, v\rangle \text{ and } \langle B, v'\rangle,\ \langle A, v\rangle \in h \text{ iff } \langle B, v'\rangle \in h.$$

Let C $(= C(x_1, \ldots, x_n))$ be a formula of $\mathscr{L}$ with exactly n free variable $x_1, \ldots, x_n$, and let v be an assignment of values to variables. We say that $\langle C, v\rangle$ *falls under* h *relative to* M (notation: $\langle C, v\rangle \in_M h$) iff there is an assignment v' and a sequence (possibly empty) of distinct names $\langle a_1, \ldots, a_m\rangle$ and a formula C' $(= C'(x_1, \ldots, x_n, y_1, \ldots y_m))$ that has precisely $n + m$ free variables, $x_1, \ldots, x_n, y_1, \ldots y_m$, and that satisfies the following conditions:

(i) $C = C'(x_1, \ldots, x_n, a_1, \ldots a_m)$;
(ii) for all i, $1 \leq i \leq n$, $v'(x_i) = v(x_i)$;
(iii) for all i, $1 \leq i \leq m$, $v'(y_i) = I(a_i)$; and
(iv) $\langle C', v'\rangle \in h$.

If $C(x)$ is a formula with exactly one free variable, then let *the extension assigned to* $C(x)$ *by hypothesis* h (notation: $h(C(x))$) be the set of those objects $d \in D$ such that

$$\exists v \in V[v(x) = d \ \&\ \langle C(x), v\rangle \in_M h].$$

Set $M + h$ to be an interpretation of $\mathscr{L}$ that is just like M except that the predicate G is assigned $h(A(x, G))$. Now, the *revision rule*, $\delta_{\mathscr{D},M}$, is an operation on hypotheses that satisfies the following condition: for all formulas B without occurrences of names and all assignments v,

$$\langle B, v\rangle \in \delta_{\mathscr{D},M}(h) \text{ iff } v \text{ satisfies } B \text{ in } M + h.$$

The revision rule provides the key semantical information about the circular predicate G. Intuitively, the rule may be understood thus: the result of applying $\delta_{\mathscr{D},M}$ to a hypothesis h (i.e., $\delta_{\mathscr{D},M}(h)$) provides a better, or at least an equally good, interpretation of G as the hypothesis h. So, if we begin with an arbitrary hypothesis, we may attempt to improve it through repeated applications, possibly transfinite, of $\delta_{\mathscr{D},M}$. The result is a *revision sequence*, which we will define after some preliminary definitions.

Let On be the class of all ordinals. Let $\mathscr{S}$ be an On-long sequence of hypotheses, and let $\mathscr{S}_\beta$ be the β^{th} member of $\mathscr{S}$. If α is a limit ordinal, then we say that $\langle B, v\rangle$ is *stably in* [*stably out of*] $\mathscr{S}$ *at* α iff

$$\exists\gamma < \alpha\forall\beta(\text{if } \gamma \leq \beta < \alpha \text{ then } \langle B, v\rangle \in [\notin]\ \mathscr{S}_\beta).$$

Similarly, we say that $\langle B, v\rangle$ is *stably in* [*stably out of*] $\mathscr{S}$ iff

$$\exists\gamma\forall\beta(\text{if } \gamma \leq \beta \text{ then } \langle B, v\rangle \in [\notin]\ \mathscr{S}_\beta).$$

Observe the following facts about arbitrary On-long sequences $\mathscr{S}$.

(i) $\mathscr{S}$ is bound to contain a *cofinal* hypothesis: a hypothesis h that occurs over and over again in $\mathscr{S}$; that is, $\forall\alpha\exists\beta \geq \alpha(h = \mathscr{S}_\beta)$.
(ii) *Reflection ordinals* are bound to exist for $\mathscr{S}$. That is, ordinals α exist at which stability coincides with stability in $\mathscr{S}$ as a whole. More precisely, at α, a pair $\langle B, v\rangle$ is stably in [stably out of] $\mathscr{S}$ at α iff $\langle B, v\rangle$ is stably in [stably out of] $\mathscr{S}$.

We say that a hypothesis h *coheres with* a sequence $\mathscr{S}$ *at* a limit ordinal α iff (i) if a pair $\langle B, v\rangle$ is stably in $\mathscr{S}$ at α then $\langle B, v\rangle \in h$, and (ii) if a pair $\langle B, v\rangle$ is stably out of $\mathscr{S}$ at α then $\langle B, v\rangle \notin h$. Finally, we say $\mathscr{S}$ is a *revision sequence for* $\delta_{\mathscr{D},M}$ iff, for all ordinals α and β,

(i) if $\alpha = \beta + 1$ then $\mathscr{S}_\alpha = \delta_{\mathscr{D},M}(\mathscr{S}_\beta)$; and
(ii) if α is limit then $\mathscr{S}_\alpha$ coheres with $\mathscr{S}$ at α.

It is easy to show that all hypotheses h generate revision sequences; that is, there is at least one revision sequence $\mathscr{S}$ such that $\mathscr{S}_0 = h$.

A hypothesis h is said to be *recurring for* a revision rule $\delta_{\mathscr{D},M}$ iff h is cofinal in a revision sequence for $\delta_{\mathscr{D},M}$. Recurring hypotheses are the survivors of the revision process, and they can be used to endow circular predicates with a logic and semantics. Gupta and Belnap provide two broad ways of doing so. Here is one way, which yields their system $\mathbf{S}^{\#}$:

(i) A sentence A is *valid relative to* $\mathscr{D}$ *in* M *in* $\mathbf{S}^{\#}$ ($M \models_{\mathscr{D}} A$) iff for all hypotheses h recurring in $\delta_{\mathscr{D},M}$ there is a natural number n such that for all natural numbers p, $n \leq p$, A is true in $M + \delta^p_{\mathscr{D},M}(h)$, where $\delta^p_{\mathscr{D},M}(h)$ is the result of p applications of $\delta_{\mathscr{D},M}$ to h.
(ii) A sentence A is *valid relative to* $\mathscr{D}$ *in* $\mathbf{S}^{\#}$ ($\models_{\mathscr{D}} A$) iff A is valid in $\mathbf{S}^{\#}$ in all models M of $\mathscr{L}^-$.
(iii) *Relative to* $\mathscr{D}$, premisses $A_1, \ldots, A_n$ *semantically entail* a conclusion B ($A_1, \ldots, A_n \models_{\mathscr{D}} B$) iff $\models_{\mathscr{D}} [(A_1 \& \ldots \& A_n) \supset B]$.

One compelling feature of $\mathbf{S}^{\#}$ is that it yields an attractive theory of *finite* definitions, definitions for which transfinite revisions are unnecessary:

> A definition $\mathscr{D}$ is *finite* iff, for all models M of $\mathscr{L}$, there is a number n such that, for all hypotheses h, $\delta^n_{\mathscr{D},M}(h)$ is finitely reflexive for $\delta_{\mathscr{D},M}$; where h' is *finitely reflexive for* $\delta_{\mathscr{D},M}$ iff, for some natural number p, $0 < p$, $h' = \delta^p_{\mathscr{D},M}(h')$.

For finite definitions $\mathscr{D}$, the notion of validity can be greatly simplified:

$M \models_{\mathscr{D}} A$ iff, for all hypotheses h finitely reflexive for $\delta_{\mathscr{D},M}$, A is true in $M + h$.

Furthermore, a simple sound and complete calculus, $\mathbf{C}_0$, can be provided for reasoning with finite definitions. In this calculus, one works with *indexed* formulas, A^i, where i is an integer. We may view the integer index as representing a stage in the revision process. The rules governing the logical connectives are classical, with the proviso that, in an application, the premisses and conclusions must be decorated with the same index. The rules for definitions, DfI and DfE, require shifts in indices, however:

DfE: $Gx^{i+1} \;/ \therefore\; A(x, G)^i$;
DfI: $A(x, G)^i \;/ \therefore\; Gx^{i+1}$.

$\mathbf{C}_0$ has one last rule, *Index Shift*, that allows arbitrary shifting of the index of a formula so long as it contains no occurrences of the defined predicate. Note finally that

$\mathbf{C}_0$ is sound for non-finite definitions also, but it fails to be complete for them.[20] For further information about $\mathbf{C}_0$ and about the revision theory of finite and other definitions, see Gupta and Belnap [34], Martinez [44], and Gupta [32].

3.2 Introducing the New Conditionals

Let us now expand our language to $\mathscr{L}^+$ by adding two new conditionals, the *step-down* conditional ($\rightarrow$) and the *step-up* conditional ($\leftarrow$). The intuitive meaning of these conditionals is roughly as follows. $(B \rightarrow C)$ says that if B is true at a revision stage then C is true at the previous stage. So, this conditional takes us a step down in the revision process (hence the name we have given it). On the other hand, $(C \leftarrow B)$ takes us a step up. It says that if B is true at the previous revision stage then C is true at the current stage.[21] Let us make these ideas precise.

Let $\mathscr{L}^-$ be, as before, a classical first-order language with interpretation M. Let $\mathscr{L}^-$ be expanded to $\mathscr{L}^+$ by adding to it the two new *step* conditionals and a one-place predicate G that is governed by the definition,

$(\mathscr{D})\quad Gx =_{Df} A(x, G).$

Note that the definiens here, $A(x, G)$, is a formula of $\mathscr{L}^+$ and thus may contain occurrences of step conditionals. Hypotheses h are defined, in a way parallel to before, as certain sets of pairs of formulas of $\mathscr{L}^+$ and assignments of values to variables. Furthermore, we define the notion "v satisfies a formula B of $\mathscr{L}^+$ relative to M and h," in symbols,

$$M, h, v \models^+ B,$$

in the familiar way, reading G as receiving the interpretation $h(A(x, G))$ but interpreting the new conditionals as follows:

$$M, h, v \models^+ (B \rightarrow C) \text{ iff either } M, h, v \not\models^+ B \text{ or } \langle C, v\rangle \in_M h; \text{ and}$$
$$M, h, v \models^+ (C \leftarrow B) \text{ iff either } \langle B, v\rangle \notin_M h \text{ or } M, h, v \models^+ C.$$

Define the *step biconditional*, $(B \leftrightarrow C)$, thus:

$$(B \rightarrow C) \,\&\, (B \leftarrow C).$$

Then, we have

$$M, h, v \models^+ (B \leftrightarrow C) \text{ iff } (M, h, v \models^+ B \text{ iff } \langle C, v\rangle \in_M h).$$

[20] In Gupta and Belnap's other system, **S***, validity receives a simpler definition: a sentence is valid iff it is true under all recurring hypotheses. However, the logic is considerably weaker, and calculus $\mathbf{C}_0$ is no longer sound.

[21] In reading these conditionals, it is a useful mnemonic to think of the left-hand side (which corresponds to the definiendum in a definition) as higher and the right-hand side (which corresponds to the definiens) as lower. So, in $(B \rightarrow C)$ we are stepping down as we move from the antecedent to the consequent, whereas in $(C \leftarrow B)$ we are stepping up. Incidentally, we read $(B \rightarrow C)$ as "B right-arrow C," "B step-down C," and also as "if B, C"; and $(C \leftarrow B)$ as "B left-arrow C," "B step-up C," and "C, if B."

The revision rule, $\Delta_{\mathscr{D},M}$, for $\mathscr{L}^{+}$ is now defined thus: for all formulas B without occurrences of names and for all assignments v,

$$\langle B, v\rangle \in \Delta_{\mathscr{D},M}(h) \text{ iff } M, h, v \models^{+} B.$$

It is a routine verification that if B is a sentence and v and v' are arbitrary assignments, then

$$M, h, v \models^{+} B \text{ iff } M, h, v' \models^{+} B.$$

So, if B is a sentence, set:

$$M, h \models^{+} B \text{ iff, for an arbitrary assignment } v, M, h, v \models^{+} B.$$

We can now adapt the earlier definitions to obtain the notions "revision sequence for $\Delta_{\mathscr{D},M}$" and "a sentence B of $\mathscr{L}^{+}$ is valid relative to $\mathscr{D}$ in M in the system $\mathbf{S}^{\#}$" ($M \models^{+}_{\mathscr{D}} B$), and thereby, the notion of entailment ($A_1, \ldots, A_n \models^{+}_{\mathscr{D}} B$). Let us illustrate these notions.

Example 3.2.1 (Definitional equivalence expressed using "$\leftrightarrow$") Let G be governed by definition $\mathscr{D}$, and let $\mathscr{S}$ be an arbitrary revision sequence for $\Delta_{\mathscr{D},M}$. Then, for all $\alpha \geq 1$,

$$M, \mathscr{S}_{\alpha} \models^{+} \forall\, x(Gx \leftrightarrow A(x, G)).$$

Hence,

$$\models^{+}_{\mathscr{D}} \forall x(Gx \leftrightarrow A(x, G)).$$

The step biconditional ($\leftrightarrow$) thus enables us to reflect the definitional equivalence governing G *within* $\mathscr{L}^{+}$. The material equivalence ($\equiv$) is inadequate for this task, for if the definiens is $\neg Gx$, then we have

$$\not\models^{+}_{\mathscr{D}} \forall x(Gx \equiv \neg Gx).$$

Nevertheless, $\forall x(Gx \leftrightarrow \neg Gx)$ remains valid. ⊣

Example 3.2.2 (Non-circular definiens from $\mathscr{L}^{-}$) Let G be defined non-circularly via the definiens $x = a$, and let a denote 0 in M. Observe that if $\mathscr{S}$ is an arbitrary revision sequence for $\Delta_{\mathscr{D},M}$, then for all $\alpha > 1$, $\mathscr{S}_{\alpha}(x = a) = \mathscr{S}_{\alpha}(Gx) = \{0\}$. Hence,

$$\begin{aligned} M \quad &\models^{+}_{\mathscr{D}} \forall x(Gx \equiv x = a); \text{ and, indeed,} \\ &\models^{+}_{\mathscr{D}} \forall x(Gx \equiv x = a). \end{aligned}$$

The expected material equivalence of the definiendum with the definiens thus holds. Suppose now that, for some assignment v, $\langle \bot, v\rangle \in \mathscr{S}_0$. Then, we have

$\langle \bot, v\rangle$ belongs to $\mathscr{S}_0$, but not to $\mathscr{S}_{\alpha}$, for any $\alpha > 0$;

$\langle(\top \rightarrow \bot)\ v$ belongs to $\mathscr{S}_1$, but not to $\mathscr{S}_{\alpha}$, for any $\alpha > 1$;

$\langle(\top \rightarrow (\top \rightarrow \bot)), v\rangle$ belongs to $\mathscr{S}_2$, but not to $\mathscr{S}_{\alpha}$, for any $\alpha > 2$; and so on.

Thus, even with non-circular definitions, the revision process can exhibit instability at least up to ordinal ω. This is due to the presence of the new conditionals. In the absence of these conditionals, the revision sequences for non-circular definitions exhibit no instability beyond the first stage.

Example 3.2.3 (Non-circular definiens from $\mathscr{L}^+$) Let the definiens of G be $(x \neq b \rightarrow x = a)$; and let a and b denote in M, respectively, 0 and 1. Now the interpretation of G in a revision sequence will eventually settle down to $\{0, 1\}$. But notice that the presence of a step conditional in the definiens entails that it may take the revision process a little longer to reach this interpretation. The greater the embedding of the step arrows within one another in a definiens, the higher in general the revision stage by which the interpretation of the definiendum is bound to become stable.

The following two theorems, which are easily verified, settle the behavior of non-circular definitions.[22]

Theorem 3.2.4 (Fixed-points of $\Delta_{\mathscr{D},M}$) *Let $\mathscr{S}$ be an arbitrary revision sequence of $\Delta_{\mathscr{D},M}$, and for some ordinal α, let $\mathscr{S}_\alpha$ be a fixed point of $\Delta_{\mathscr{D},M}$—that is, let $\Delta_{\mathscr{D},M}(\mathscr{S}_\alpha) = \mathscr{S}_\alpha$. Then (a) for all ordinals $\beta \geq \alpha$, $\mathscr{S}_\alpha = \mathscr{S}_\beta$; and (b) for all formulas B and C and all assignments v, the following are equivalent:*

(i) $M, \mathscr{S}_\alpha, v \models^+ (B \rightarrow C)$,
(ii) $M, \mathscr{S}_\alpha, v \models^+ (C \leftarrow B)$, *and*
(iii) $M, \mathscr{S}_\alpha, v \models^+ (B \supset C)$.

Theorem 3.2.5 (Non-circular definitions) *Let the definiens of G be a formula $A(x)$ of $\mathscr{L}^+$ that contains no occurrences of G. Let n be the number of occurrences of the step arrows in $A(x)$, and let $\mathscr{S}$ and $\mathscr{S}'$ be arbitrary revision sequences for $\Delta_{\mathscr{D},M}$. Then:*

(i) *for all ordinals $\alpha > n + 1$, $\mathscr{S}_\alpha(Gx) = \mathscr{S}'_{n+2}(Gx)$; and*
(ii) *for all ordinals $\alpha \geq \omega$, $\mathscr{S}_\alpha = \mathscr{S}'_\omega$.*

Hence, all revision sequences for $\Delta_{\mathscr{D},M}$ culminate in the same fixed point, and the distinction between the step conditionals and the material conditional collapses. We have:

(iii) $\models^+_{\mathscr{D}} \forall x(Gx \equiv A(x))$.

Example 3.2.6 (Circular definiens from $\mathscr{L}^+$) Let the definiens of G be $(Gx \rightarrow Gx)$, and let M be an arbitrary ground model with domain D. Further, let X and Y be arbitrary subsets of D, and let $\mathscr{S}$ be an arbitrary revision sequence for $\Delta_{\mathscr{D},M}$ such that

$$\mathscr{S}_0(Gx) = X \text{ and } \mathscr{S}_0(Gx \rightarrow Gx) = Y.$$

[22] In the interest of brevity, we omit proofs in this essay. The reader will find proofs as well as a more detailed presentation of the theory in Standefer [66, 68].

Then it can be verified that, for all $n \geq 0$,

$$\mathscr{S}_{2n+1}(Gx) = Y \text{ and } \mathscr{S}_{2n+2}(Gx) = X \cup (D \setminus Y).$$

Observe that the resulting revision pattern for G is different from the one we obtain with the definiens $(Gx \supset Gx)$: in the latter revisions, G invariably receives the entire domain, D, as its interpretation. More generally, the revision patterns we obtain with the definiens $(Gx \rightarrow Gx)$ are essentially new. These patterns do not occur if G is defined using only the resources of $\mathscr{L}$ (i.e., without the step conditionals). The addition of the step conditionals makes, therefore, an essential difference to the revision environment.

The hypotheses of revision are infinite sets. The next theorem shows that only a finite part of this set is relevant to the subsequent behavior of a formula in a revision sequence.

Theorem 3.2.7 (Revision sequences for circular definiens from $\mathscr{L}^+$) *Let the definiens of G be an arbitrary formula $A(x, G)$ of $\mathscr{L}^+$, and let $\mathscr{S}$ and $\mathscr{S}'$ be arbitrary revision sequences for $\Delta_{\mathscr{D},M}$ such that, for all subformulas B of $A(x, G)$,*

$$\mathscr{S}_0(B) = \mathscr{S}'_0(B).$$

Then, if C is an arbitrary formula with n occurrences of the step conditionals, then, for all m, $m > n$: $\mathscr{S}_m(C) = \mathscr{S}'_m(C)$.[23]

3.3 The Logic of the Step Conditionals

Let us begin by observing that the logical behavior of the step conditionals is highly unusual. The following fails to be a logical law:

$$\neg(A \leftrightarrow \neg A).$$

Relatedly, Identity is not a logical law for the step-down conditional. That is, for some definitions $\mathscr{D}$,

$$\not\models^+_{\mathscr{D}} A \rightarrow A.$$

A parallel claim holds for the step-up conditional. Notice also that the two conditionals are governed by different logics. Thus, Exportation holds for the step-up conditional but fails for the step-down conditional:

$$\models^+_{\mathscr{D}} (C \leftarrow A \;\&\; B) \supset ((C \leftarrow B) \leftarrow A);$$
$$\not\models^+_{\mathscr{D}} (A \;\&\; B \rightarrow C) \supset (A \rightarrow (B \rightarrow C)).$$

All the invalidities above can be verified by letting $\mathscr{D}$ be a simple circular definition—for example, $Gx =_{Df} \neg Gx$.

We shall argue below that the step conditionals are the right ones for formulating the rules for truth, (TI) and (TE). We wish to notice now that, although the logical behavior of the step conditionals is unusual, it is simple and straightforward. Let

[23] Note that every formula counts as a subformula of itself.

us observe, first, that the step conditionals have natural introduction and elimination rules. The elimination rule for the step-down conditional allows us to infer the indexed formula C^i from $(B \to C)^{i+1}$ and B^{i+1}. The introduction rule allows us to conclude $(B \to C)^{i+1}$ if we can derive C^i from B^{i+1}. The rules for the step-up conditional are parallel. In a Fitch-style deduction system, these rules can be represented as follows.

$$\begin{array}{|ll} \begin{array}{|l} B^{i+1} \\ \hline \\ \vdots \\ C^{i} \end{array} & \begin{array}{l} \text{hyp} \\ \\ \\ \\ \end{array} \\ (B \to C)^{i+1} & \to\text{I} \end{array} \qquad \begin{array}{|ll} (B \to C)^{i+1} & \\ \vdots & \\ B^{i+1} & \\ \vdots & \\ C^{i} & \to\text{E} \end{array}$$

$$\begin{array}{|ll} \begin{array}{|l} B^{i} \\ \hline \\ \vdots \\ C^{i+1} \end{array} & \begin{array}{l} \text{hyp} \\ \\ \\ \\ \end{array} \\ (C \leftarrow B)^{i+1} & \leftarrow\text{I} \end{array} \qquad \begin{array}{|ll} (C \leftarrow B)^{i+1} & \\ \vdots & \\ B^{i} & \\ \vdots & \\ C^{i+1} & \leftarrow\text{E} \end{array}$$

Let $\mathbf{C}_0^+$ be $\mathbf{C}_0$ with the addition of the introduction and elimination rules for both step conditionals. The next two theorems relate the calculus $\mathbf{C}_0^+$ to validity in $\mathscr{L}^+$.

Theorem 3.3.1 (Soundness of $\mathbf{C}_0^+$) *For every sentence B of $\mathscr{L}^+$, if B is a deducible in $\mathbf{C}_0^+$ from definition $\mathscr{D}$ (notation: $\vdash_{\mathscr{D}}^{+} B$) then B is valid in $\mathscr{L}^+$ relative to $\mathscr{D}$ (i.e., $\models_{\mathscr{D}}^{+} B$).*

Theorem 3.3.2 (Soundness and completeness of $\mathbf{C}_0^+$ for finite definitions) *Let $\mathscr{D}$ be a definition without any occurrences of the step conditionals. If $\mathscr{D}$ is a finite definition of $\mathscr{L}$, then for all sentences B of $\mathscr{L}^+$: $\vdash_{\mathscr{D}}^{+} B$ iff $\models_{\mathscr{D}}^{+} B$.*[24,25]

[24] Because of the presence of step conditionals in $\mathscr{L}^+$, no definition—not even a noncircular one—yields finitely reflexive hypotheses. (See Example 3.2.2.) Hence, no definition meets the requirement for finiteness, as this requirement is set out in Section 3.1. We can, by liberalizing the requirement, obtain a notion of finite definition suitable for $\mathscr{L}^+$ (see Standefer [68]). Here, however, we have chosen to restrict the theorem to those definitions formulated in $\mathscr{L}$ which count as finite definitions *of* $\mathscr{L}$.

[25] The full logic of circular definitions is highly complex (in the recursion-theoretic sense) and is not axiomatizable; see Kremer [38], Antonelli [2, 3], and Kühnburger *et al.* [40]. The complexity of revision theories of truth has been studied by, among others, Burgess [14], Löwe and Welch [42], and Welch [74].

$\mathbf{C}_0^+$ is not, in general, complete with respect to validity in $\mathbf{S}^\#$. There is, however, a weaker sense of validity for which it is complete, regardless of the set of circular definitions.[26].

Let us observe, next, that the two step-conditionals are interdefinable. The following equivalences hold:

$$(B \leftarrow A) \equiv ((\top \rightarrow A) \supset B), \text{ and}$$
$$(A \rightarrow B) \equiv (A \supset (\bot \leftarrow \neg B)).$$

Moreover, since the following is a logical law,

$$(\top \rightarrow A) \equiv (\bot \leftarrow \neg A),$$

each of the step conditionals is interdefinable with a modality (□). For instance, we can define □ in terms of → using the following equivalence:

P□: $\Box A \equiv (\top \rightarrow A)$.

Or, alternatively, we can define → in terms of □ using the following principle:

P→: $(A \rightarrow B) \equiv (A \supset \Box B)$.

A parallel principle enables us to define ← directly in terms of □:

P←: $(B \leftarrow A) \equiv (\Box A \supset B)$.

Observe, finally, that □ is governed by a simple normal modal logic. If we set out this logic, we can recover the logic of the step conditionals from it.

Let **SC**, the logic of the step conditionals, be the axiomatic system consisting of (i) a set of axioms for classical first-order logic with identity, as set out in, say, Mendelson [49]; (ii) P□, which we treat as a definition of □; (iii) P→ and P←; and (iv) for all formulas A and B, and all G-free formulas C, of $\mathscr{L}^+$, the following formulas:

$$\Box(A \supset B) \supset (\Box A \supset \Box B),$$
$$\neg\Box\neg A \equiv \Box A,$$
$$\forall x \Box A \equiv \Box \forall x A, \text{ and}$$
$$C \equiv \Box C.$$

The rules for inference for **SC** are (a) modus ponens (from $A \supset B$ and A to infer B); (b) universal generalization (from A to infer $\forall x A$); and (c) necessitation (from A to infer $\Box A$). We understand the notion "theorem of **SC**" ($\vdash_{sc}$) in the usual way. Now one can establish the following theorem, which connects the calculus **SC** with the revision semantics given above.

[26] The weaker sense is "validity in $\mathbf{S}_0$"; for the definition of this notion, see Gupta and Belnap [34, p.147]

Theorem 3.3.3 *[Soundness and completeness for* **SC***] For all sentences A of $\mathscr{L}^+$: $\vdash_{sc} A$ iff, for all definitions $\mathscr{D}$, $\models^+_{\mathscr{D}} A$.* [27]

If we drop the introduction and elimination rules for definitions from $\mathbf{C}_\mathbf{0}^+$, then the resulting system—**FSC**, to give it a name—is equivalent to the system **SC**:

Theorem 3.4.1 (Equivalence of **SC** with **FSC**) *For every sentence B of $\mathscr{L}^+$, B is a theorem of* **SC** *iff B is a theorem of* **FSC**.

3.4 An Informal Account of the Step Conditionals

A definition warrants two distinguishable logical transitions: first, from the definiendum to the definiens; and second, from the definiens to the definiendum. For non-circular definitions in, say, a classical context, the logic of these transitions is captured, at least in part, by the material conditional ($\supset$). Not so, however, once we allow circular and interdependent definitions. It is this consideration that motivates the introduction of the step conditionals. The step conditionals serve the role for definitions *in general* that the material conditional serves for ordinary, non-circular definitions.[28]

The step-down conditional ($\rightarrow$) allows us to represent the move from the definiendum to the definiens; the step-up conditional ($\leftarrow$), the move from the definiens to the definiendum. If, as in revision semantics, we think of the interpretation of the definiendum as occupying a stage higher than the corresponding interpretation of the definiens, then the step-down conditional takes us from the higher definiendum stage to the lower definiens stage and, analogously, the step-up conditional takes us from the lower definiens stage to the higher definiendum stage. With non-circular definitions, the distinction between stages is inessential and, hence, so also is the distinction between the two conditionals: both step conditionals reduce to the material conditional. With circular definitions, however, the distinction between stages and conditionals is vital.

The precise semantics for our two conditionals, spelled out above, is complex because not only do we use these conditionals to represent facts about circularly defined predicates, we allow these conditionals to occur in the definientia of new circular definitions. This creates technical complications, and we need to proceed with care. Still, the intuitive idea behind the conditionals is straightforward. Consider a revision sequence of better and better evaluations of the language. Then, $(A \rightarrow B)$ is true at a stage $\alpha + 1$ iff, if A is true at $\alpha + 1$, then B is true at α—here we are stepping down to the conclusion. And, $(B \leftarrow A)$ is true at $\alpha + 1$ iff, if A is true at α, then B is true at $\alpha + 1$—here we are stepping up to the conclusion.

It is an easy consequence of the semantics of the step conditionals that any definition, $\mathscr{D}$,

[27] See Standefer [68].

[28] Here and in the rest of the present subsection, we assume that the ground language is classical.

$(\mathscr{D})\quad Gx =_{Df} A(x, G),$

whether circular or not, implies the conditionals,

$$Gx \rightarrow A(x, G);\ \text{and}$$
$$Gx \leftarrow A(x, G).$$

Let us define the step biconditional, $(B \leftrightarrow C)$, thus:

$$(B \rightarrow C)\ \&\ (B \leftarrow C).$$

Then, definition $\mathscr{D}$ invariably implies

$$\forall x(Gx \leftrightarrow A(x, G)),$$

although $\mathscr{D}$ does not, in general, imply

$$\forall x(Gx \equiv A(x, G)).$$

With circular definitions, it is essential to distinguish between the material biconditional and the step biconditional. Only thus can we sustain the idea of coherent circularity.

The step conditionals are far from ordinary, familiar conditionals (such as strict conditionals and counterfactual conditionals). For example, reflexivity fails for them: neither $(A \rightarrow A)$ nor $(A \leftarrow A)$ is a logical law. For another example, $(A \leftrightarrow B)$ is not symmetric; the following fails to be a logical law:

$$(A \leftrightarrow B) \equiv (B \leftrightarrow A).$$

We wish to make three observations to mitigate the seeming counterintuitiveness of these results.

(i) The non-standard character of the conditionals issues directly from the function they are designed to serve. If, for example, we require either one of $(A \rightarrow A)$ and $(A \leftarrow A)$ to be a logical law, then the distinction between these conditionals and the material conditional collapses, and we cannot make sense of coherent circularity.

(ii) The non-standard behavior does not undermine the idea that our connectives capture readings of 'if'. The general notion expressed by the English 'if' is that *if* the antecedent is true at a certain point of evaluation e *then* the consequent *must* be true at a related point of evaluation e^*. We obtain specific readings of 'if' by understanding the force of the modality "must" in a particular way and by imposing specific relations between the evaluation points e and e^*. Thus, in the material conditional, the force of 'must' is entirely negated and the evaluation points are identified. In strict conditionals, the evaluation points are identified but the modality retains a significant force.[29] Conditionals used in connection with generics ("if you strike a match, it lights") depart from the

[29] The same is true of David Lewis's counterfactual conditionals, though with these conditionals the force of 'must' depends in part on the antecedent. The entailment conditionals studied by Alan Ross Anderson, Belnap, and others also retain modal force. On the other hand, the conditional defended by Robert Stalnaker lacks modal force.

material conditional still further: in them a strong modality remains in force, and furthermore, the consequent is evaluated at points that are, in general, different from those of the antecedent. (In the example given, the consequent is evaluated at a moment later than that at which the antecedent is evaluated.) Our step conditionals are like the material conditional in that in them, too, the modality carries no force, but they are like the 'if' used with generics in that the evaluation points for the antecedent and the consequent can be distinct. As a result, reflexivity fails for the step conditionals, but conditional excluded middle holds. The following schemata are logically valid:

$$(A \rightarrow B) \vee (A \rightarrow \neg B); \text{ and}$$
$$(B \leftarrow A) \vee (\neg B \leftarrow A).$$

The logic of the English 'if' is not simple, and this logic is not stable across different uses of 'if'. In one set of its uses, namely, those where 'if' is used to express the relationship between definienda and definientia, its logic is captured, we believe, by the step conditionals.

(iii) The basic logic of the step conditionals, though non-standard, is quite simple—simpler, we think, than that of other non-classical conditionals. We offered in the previous section simple calculi for reasoning with these conditionals, including a Fitch-style natural deduction system for the step conditionals. At an informal level, perhaps the observations that follow will bring out the simplicity of the logic.

Let us introduce a connective ($\Box$) with the following semantics: $\Box A$ is true at a stage $\alpha + 1$ iff A is true at α.[30] So, intuitively, $\Box A$ says that A is true at the previous revision stage. The meaning of the two conditionals can now be captured through the following equivalences:

P→: $(A \rightarrow B) \equiv (A \supset \Box B)$; and
P←: $(B \leftarrow A) \equiv (\Box A \supset B)$.

Let us notice, next, that $\Box$ is governed by a normal modal logic in which the following schemata are logically valid:

$$\Box(A \vee B) \equiv (\Box A \vee \Box B),$$
$$\Box\neg A \equiv \neg\Box A, \text{ and}$$
$$\forall x \Box A \equiv \Box \forall x A.$$

Thus, $\Box$ can be distributed across any logical connective in a formula, and the result is a logically equivalent formula. In particular, $\Box$ can be pushed into the interior of the formula so that no other logical connective lies within its scope. Let the *normal form* of a formula be the formula that results when (a) the step conditionals are eliminated using P→ and P←, (b) all the occurrences of the $\Box$ are pushed as far as possible into the interior of the formula, and (c) all occurrences of the $\Box$ next to identities are

[30] Note that $\Box$ carries no modal force. We use this symbol because many of the laws governing it have been studied in modal logic.

deleted. For example, let us understand A, B, and C to be atomic in the formula

$$(C \leftarrow A \,\&\, B) \supset ((C \leftarrow B) \leftarrow A).$$

Then, step (a) applied to this formula yields

$$(\Box(A \,\&\, B) \supset C) \supset (\Box A \supset (\Box B \supset C));$$

and an application of step (b) now yields

$$((\Box A \,\&\, \Box B) \supset C) \supset (\Box A \supset (\Box B \supset C)),$$

which is the normal form of the initial formula. Let a normal form be *classically valid* iff it is a theorem of classical logic when all occurrences of $\Box$ formulas are taken to be atomic.[31] Then, the normal form just displayed is classically valid, and this suffices to establish, by the following theorem, the logical validity of the original formula, in the sense of the normal modal logic indicated above. Exportation is thus valid for the step-up conditional, though, as we saw above, it fails for the step-down conditional.

Theorem 3.4.1 (An equivalence concerning normal forms) *A formula containing step conditionals is logically valid iff its normal form is classically valid.*

It follows immediately that the propositional logic of the step conditionals is decidable and, more generally, that the question of the validity of any step formula reduces to a question about classical validity.

We can now easily verify that the following schemata are valid. (The first two of these schemata reflect the classicality of the ground language.)

$$(A \rightarrow \neg\neg B) \supset (A \rightarrow B)$$
$$((A \rightarrow B) \,\&\, (A \rightarrow \neg B)) \supset \neg A$$
$$((A \rightarrow C) \,\&\, (B \rightarrow C)) \supset ((A \vee B) \rightarrow C)$$
$$((A \vee B) \rightarrow C) \supset ((A \rightarrow C) \,\&\, (B \rightarrow C))$$
$$\forall x(A \rightarrow B) \supset (\forall x A \rightarrow \forall x B)$$
$$((A' \supset A) \,\&\, \Box(B \supset B') \,\&\, (A \rightarrow B)) \supset (A' \rightarrow B')$$

Note that these schemata are not valid if occurrences of the material conditional are replaced by either of the step conditionals. Note also that analogues of the above schemata hold also for the step-up conditional.

Define by recursion:

$$\Box^0 A = A; \text{ and}$$
$$\Box^{n+1} A = \Box\Box^n A.$$

And set:

$$(A \leftrightarrow_n B) = (A \equiv \Box^n B).$$

[31] For example, all occurrences of $\Box\Box Fxy$ could be read as Gxy and all occurrences of $\Box Fxy$ as Hxy.

Then, for all natural numbers n, the following schemata are valid:

$$(A \supset \Box^n \bot) \supset \neg A, \text{ and } \Box(A \leftrightarrow_n B) \equiv (\Box A \leftrightarrow_{n+1} B).$$

Finally, let us note some schemata that fail to be valid:

$$A \rightarrow A;$$
$$(\neg A \rightarrow \neg B) \supset (B \rightarrow A);$$
$$(A \leftrightarrow_n B) \equiv (A \leftrightarrow_m B), \text{ for } m \neq n; \text{ and}$$
$$\forall x A \rightarrow A(t/x), \text{ where } t \text{ is free for } x \text{ in } A.$$

The logic of the step conditionals is definitely unusual. However, its propositional fragment is simple, and taken as a whole, the logic is no more complex than classical logic.

A Comparison In his Hilbert-Style axiomatization of calculi for circular definitions, Bruni [11] also extends the language with a new conditional. The principal differences between Bruni's conditional and the conditionals introduced here are these: (i) Bruni's conditional connects indexed formulas whereas our step conditionals connect formulas *simpliciter*. (ii) Bruni's conditional does not belong to the language in which circular definitions are formulated but, as Bruni puts it, the conditional is "superimposed." Consequently, Bruni's conditional cannot occur in the definiens of circular definitions. Our step conditionals, in contrast, may be used to construct new circular definitions. For example, the definition

$$Gx =_{Df} Gx \leftrightarrow \neg Gx$$

can be treated within the expanded revision theory, but the corresponding definition that uses Bruni's conditional cannot be treated by Bruni's calculi.

We think that the step conditionals makes available new, and perhaps simpler, versions of Bruni's sequent calculi for circular definitions.

3.5 Application to Truth

The above account of definitions and conditionals yields a theory of truth if we read the *Tarski biconditionals*,

$$T(\text{`}A\text{'}) \text{ iff } A,$$

as partial definitions of truth. Under the assumption that only sentences are true, the partial definitions fix a rule of revision, and the semantical scheme sketched above applies. Theorem 3.3.1 continues to hold with the extended notion of definition: $\mathbf{C}_\mathbf{0}^+$ is a sound calculus for truth and the step conditionals.

This theory of truth enables us to express (TI) and (TE) in the object language. We propose that the 'if' in (TI) should be read as the step-up conditional, and the 'if' in (TE) as the step-down conditional. (TI) and (TE) may now be expressed thus:

(TI$^\text{s}$) $T(\text{`A'}) \leftarrow A$; and
(TE$^\text{s}$) $T(\text{`A'}) \rightarrow A$.

Consequently, the Tarski biconditionals can also be expressed in the object language. The 'iff' in

$$T(\text{‘}A\text{’}) \text{ iff } A,$$

should, we propose, be read as the step biconditional. That is, we should understand the Tarski biconditionals thus:

$$T(\text{‘}A\text{’}) \leftrightarrow A.$$

We shall call this the *step T-biconditional for A*. The *material T-biconditional for A*, in contrast, is this equivalence:

$$T(\text{‘}A\text{’}) \equiv A.$$

We now argue that the resulting theory of truth satisfies the desiderata laid down in Section 1.

3.5.1 Descriptive Adequacy

We make the following observations in favor of the descriptive adequacy of the resulting theory:

(i) *Tarski biconditionals.* The proposed reading of 'iff' allows us to unrestrictedly accept the Tarski biconditionals and, thus, respect a motivation for the inconsistency view (Section 2.1). We accept, for example, the Tarski biconditional for the Liar sentence l (where $l = \text{‘}\neg Tl\text{’}$):

$$T(\text{‘}\neg Tl\text{’}) \leftrightarrow \neg Tl.$$

But this biconditional implies no contradictions. In contrast, under the material reading favored by the inconsistency view, namely,

$$T(\text{‘}\neg Tl\text{’}) \equiv \neg Tl,$$

the Tarski biconditional for the Liar sentence does imply a contradiction.[32] There is, thus, a vast difference between the two ways of reading 'iff'. Under one reading, the Tarski biconditionals express a law governing the concept of truth; under the other reading, they do not.[33] Indeed, under the material reading some of the Tarski biconditionals are false, as are some of the instances of (TI) and (TE). Neither of the following is a law governing truth:

(TI$^{\supset}$) $A \supset T(\text{‘}A\text{’})$, and

[32] Here, and in the rest of this subsection, we assume that the ground language is classical.

[33] We note that a distinction between two readings of 'iff' is also drawn in earlier discussions of revision theory. (See, Gupta and Belnap [34, p. 138].) The distinction drawn there is that between reading 'iff' as material equivalence and reading it as definitional equivalence. This distinction is certainly a significant one. It is, however, insufficient for the purposes of the theory of truth. It is insufficient to provide, for example, an interpretation under which

$$T(\text{‘}T\text{’}(\text{‘}A\text{’})) \text{ iff } A$$

is true. Neither the material reading of 'iff' nor the definitional one yields an interpretation that is true. For further examples, see (vi) below.

(TE$^{\supset}$) $T(\text{‘}A\text{’}) \supset A$.

Neither provides a basis for a good classical theory of truth.

(ii) *The meaning of truth.* Under the proposed reading, the step biconditionals fix the meaning of truth, in the following sense: given any ground model, the biconditionals fix the revision rule for the truth predicate (assuming, as before, that only sentences are true). The theory thus preserves the strong intuitive linkage between the Tarski biconditionals and the concept of truth.[34]

(iii) *Paradoxes.* The theory rules that the reasoning in Curry's Paradox, (A7), is invalid.

(A7) Given: b = ‘if b is true then God exists’. Suppose, b is true. So, if b is true, then God exists. Hence, God exists. We can conclude, therefore, that if b is true then God exists. Hence, b must be true. So, God exists.

The transition in the very first step—from “b is true” to “if b is true, then God exists”—is improper. In the theory we have offered, the rules of inference

(TI^{R}) A; therefore, $T(\text{‘}A\text{’})$, and
(TE^{R}) $T(\text{‘}A\text{’})$; therefore, A

are *admissible*, in the sense that they hold unrestrictedly in categorical contexts: if A is categorically asserted, then $T(\text{‘}A\text{’})$ may be categorically asserted; and conversely. However, in a hypothetical context, as in the first step in Curry's Paradox, unrestricted application is illegitimate. Note that the following inference rules are unrestrictedly valid:

(TI$^{\Box}$) $\Box A$; therefore, $T(\text{‘}A\text{’})$, and
(TE$^{\Box}$) $T(\text{‘}A\text{’})$; therefore, $\Box A$.

These rules are insufficient, however, to generate Curry's Paradox. From the assumption that b is true, we can deduce □(if b is true then God exists) and, then, if □(b is true) then God exists. But this does not allow us to move to the conclusion that God exists, for we cannot deduce □(b is true) from the initial assumption. The rule “A; therefore, $\Box A$” is not a valid inference rule.

A similar analysis shows that the argument in the Liar reasoning (A6) is invalid. Furthermore, the rules governing truth allow no deduction of ‘snow is black’ from the Truth-Teller; (A10) is also invalid. Indeed, the semantics sketched above provides an explanation of *why* (A10) is invalid.

(iv) *Non-vicious reference.* If a sentence, say A, does not contain the truth predicate then its step T-biconditional implies its material T-biconditional. For now we have that

$$A \equiv \Box A.$$

The same holds for many sentences containing the truth predicate so long as they do not involve vicious reference (such as that found in the Liar). An

[34] Note that this linkage is not strong enough to sustain a deflationism about truth; see Gupta [30, 31]. For further discussion of deflationism in the context of revision theory, see Restall [58] and Yaqūb [80].

example is the sentence Ta, where a denotes

$$(T(\text{‘}Ta\text{’}) \ \&\ T(\text{‘}\neg Ta\text{’})).$$

The distinction between the two readings of the Tarski biconditional, though essential, can thus be neglected when no vicious reference is in play.[35]

(v) *Ground logic.* The logic of the ground language is not disturbed. The theory allows us to reason with ¬, &, ∀, etc. in classical ways even in the presence of vicious self-reference. Arguments (A3), (A8), and (A9) are ruled valid by the theory. The validity of these arguments, as well as that of many others, can be established in the calculi defined above.[36]

(vi) *Expressive power.* The theory provides readings under which (A4) and (A5) are valid. Argument (A4) is valid if the conditional in its conclusion is read as the step-up conditional; (A5), on the other hand, is valid if the conditional in its conclusion is read as the step-down conditional. None of these readings can be captured in classical revision theory, which provides only the invalid material conditional readings. Another example that illustrates the expressive weakness of the classical theory is this:

$$(T(\text{‘}\neg Tl\text{’}) \text{ iff } \neg Tl) \ \&\ Q.$$

Here Q is a truth and l denotes, as before, $\neg Tl$. Plainly, there is an interpretation of this sentence under which it is true. However, ‘iff’ cannot be read here as definitional equivalence, for this kind of equivalence cannot be embedded within truth-functional constructions (at least in classical revision theory). Furthermore, under the material equivalence reading of ‘iff’ the sentence is false. In contrast, if we read ‘iff’ as expressing the step biconditional, we gain an interpretation under which the sentence is true.

(vii) *Iterated truth.* Define by recursion:

$$T^0(\text{‘}A\text{’}) = A; \text{ and } T^{n+1}(\text{‘}A\text{’}) = T(\text{‘}T^n(\text{‘}A\text{’})\text{’}).$$

[35] Revision theorists have claimed it as a virtue of their theory that truth behaves like a classical concept when there is no vicious reference in the language. For a helpful discussion of this claim, see Kremer [39] and Wintein [76]. We note that we are unable to accept the desideratum Kremer titles “MGBD” on p. 357 of his essay.

[36] At the suggestion of an anonymous referee, we provide a couple of illustrations. Consider, first, argument (A9), which we may formalize thus:

$$t = \text{‘}Tt\text{’}, l = \text{‘}\neg Tl\text{’}; \text{ therefore, } \neg Tl \supset (Tl \supset Tt).$$

To establish the validity of this argument in $\mathbf{C}_\mathbf{0}^+$, it suffices to derive $(\neg Tl \supset (Tl \supset Tt))^0$ from $(t = \text{‘Tt’})^0$ and $(l = \text{‘}\neg Tl\text{’})^0$. So, suppose $\neg Tl^0$ and derive $(Tl \supset Tt)^0$. To do this, suppose Tl^0 and derive Tt^0. Now, the availability of classical rules for formulas with the same index enables us to complete the derivation.

Consider, next, argument (A4), which we may formalize thus:

$$S; \text{ therefore, } T(\text{‘}E \ \&\ S\text{’}) \leftarrow E.$$

To establish validity, suppose S^0, and derive $(T(\text{‘}E \ \&\ S\text{’}) \leftarrow E)^0$. By ←I, it suffices to to suppose E^{-1} and derive $T(\text{‘}E \ \&\ S\text{’})^0$. Since S is T-free, by Index Shift, we obtain S^{-1}. So, by & I, $(E \ \&\ S)^{-1}$. The partial definitions governing truth now yield the desired $T(\text{‘}E \ \&\ S\text{’})^0$.

Then, for all natural numbers n, the following is a logical truth:

$$T^{n}(‘(T(‘A’) \leftrightarrow A)’).$$

We can assert of a Tarski biconditional that it is true, that the truth attribution to it is true, and so on. Note that the following is not a logical truth:

$$(T^{n}(‘A’) \leftrightarrow A).$$

The biconditional governing iterated truth is properly formulated thus:

$$(T^{n}(‘A’) \leftrightarrow_{n} A).$$

More generally, the following is a law governing the truth predicate:

(IT) $\quad (T^{m+n}(‘A’) \leftrightarrow_{m} T^{n}(‘A’)).$

But the following is not, in general, valid:

$$(T^{m+n}(‘A’) \equiv T^{n}(‘A’)).$$

Let us notice that the step connectives provide an essential resource for the expression of laws governing truth. An explanation of this is as follows: Wherever circular (and interdependent) concepts are in play, the distinction between revision stages is of vital importance. Some of the laws governing circular concepts concern what happens across stages; others concern what happens within a stage. The usual logical resources (e.g., the truth-functional connectives) are adequate for the expression of *intra-stage* laws, but they are not adequate for the expression of *inter-stage* laws. The expression of the latter laws requires step connectives. $(\mathrm{TI}^{\mathrm{s}})$, $(\mathrm{TE}^{\mathrm{s}})$ and (IT) are examples of inter-stage laws. These laws require the step connectives for their proper formulation.[37]

[37] The step conditionals make available new axiomatic theories of truth that deserve study. Let us take note of three such theories. The background logic of all these theories is classical logic supplemented with the logical rules for the step conditionals.

(i) The basic theory, $\mathbf{T}_0$, consists of all the step T-biconditionals.
(ii) The next theory, $\mathbf{T}_1^{\mathrm{BG}}$, is richer. It is formulated in an extended language that allows quantification in and out of quotes (see Belnap and Grover [9]), and it consists of all the step T-biconditionals together with the generalization that says that all these biconditionals are true.
(iii) The final theory, $\mathbf{T}_2^{\mathrm{PA}}$, is formulated within arithmetized syntax and consists of the step T-biconditionals and the generalization that says that all instances of (IT) are true.

These theories are of special interest because they are different expressions of the idea that the Tarski biconditionals are the fundamental principles governing truth. Furthermore, unlike many axiomatic theories of truth considered in the literature, these theories preserve the generalization function of truth in expressively rich languages (see Section 4.3 below).

For axiomatic theories of truth, see Friedman and Sheard [24], an early and important investigation. See also Halbach [35] and Horsten [37], which provide accessible and illuminating accounts of these theories. Peter Aczel, Andrea Cantini, and Solomon Feferman did pioneering work on axiomatic theories. See Cantini [15] for references and for a masterful overview of one-hundred years of work on the paradoxes; see also Section 3.6 below.

(viii) *Semantic laws.* Laws of semantic composition are examples of intra-stage laws; they need to be formulated using the material conditional and biconditional. Examples of these laws are:

$$T(\text{`}\neg A\text{'}) \equiv \neg T(\text{`}A\text{'}),$$
$$T(\text{`}A \mathbin{\&} B\text{'}) \equiv (T(\text{`}A\text{'}) \mathbin{\&} T(\text{`}B\text{'})), \text{ and}$$
$$T(\text{`}\forall x Gx\text{'}) \supset T(\text{`}\ Gb\text{'}).$$

In the above, schematic, formulation, these laws are implied by the step T-biconditionals and are, therefore, validated by the theory we offer.[38] Note that if we replace the material conditional and biconditional in these laws by their step counterparts, the results are not in general valid. Versions of semantic laws hold also for iterated truth. For example, we have:

$$T^n(\text{`}A \mathbin{\&} B\text{'}) \equiv (T^n(\text{`}A\text{'}) \mathbin{\&} T^n(\text{`}B\text{'})).$$

Let us consider a possible objection that may be directed against us. “Your theory,” an objector may say to us, “provides a reading of ‘iff’ that preserves the Tarski biconditionals. However, your theory fails to preserve the intuition that the following variant of the biconditionals also holds:

(∗) A iff $T(\text{`}A\text{'})$.

Formula (∗) is not valid if we read ‘iff’ as material equivalence or as your step biconditional. You fail, therefore, to preserve an important intuition.”

Response. We have pointed out two different readings of ‘iff’, readings that have escaped notice before. Intuitions about what one is inclined to say when one is confused about the readings of ‘iff’ should be treated with care; we cannot rely on them uncritically. Now, we can easily provide a reading of ‘iff’ on which (∗) is valid. However, instead of multiplying readings of ‘iff’, it is better to recognize that the intuition that (∗) is valid is a product of a confusion of two readings of ‘iff’: ‘iff’ as material biconditional and as step biconditional. The confusion leads one to suppose that the step biconditional is governed by a logic more suited to the material biconditional. It leads to the idea that the ‘iff’ in Tarski biconditionals is symmetric, and thus to the thought that (∗) is valid. We are better off rejecting the confusion and along with it (∗).

Note that several variants of the above objection are possible, all based on the mismatch between the logical behavior of the step biconditional and intuitions about

[38] The step T-biconditionals do not imply generalized versions of these laws (e.g., “a conjunction is true iff its conjuncts are true”). They fail to imply also the semantic principle about universal generalization used in McGee’s ω-inconsistency theorem. It is a matter of fundamental disagreement among theorists of truth whether semantic laws should be accepted even at the cost of ω-inconsistency. We ourselves side in favor of the semantic laws. The point we wish to stress, however, is that one can endorse the step T-biconditionals without endorsing the semantic laws that generate ω-inconsistency. The readings we are offering of the conditional are neutral on how one responds to McGee’s theorem.

the logic of 'iff'. Our responses to all these objections would parallel what we have just said.[39]

The above considerations, though not conclusive, provide some evidence for the descriptive adequacy of the theory we offer. The literature on truth and the semantic paradoxes is rich, and it certainly provides theories that can claim several of the features listed above. However, as far as we know, none of the available theories can claim *all* the features.

3.5.2 Generality Requirement

The method presented above of enriching a language with circular definitions and step conditionals is highly general. Our presentation assumed, it is true, that the ground language is classical; but the assumption is inessential. The method can be applied even when the ground language is governed by a non-classical logic—such as a many-valued or a relevance or intuitionistic logic. In all cases, we obtain a theory of truth that validates Aristotle's Rules, interpreted, as above, using the step conditionals. Indeed, thus interpreted, Aristotle's Rules fix, irrespective of the ground logic, the revision rule governing truth. Thus, in one sense of meaning, the Rules fix the meaning of the truth predicate. More generally, theories of truth we gain for non-classical languages possess virtues similar to those indicated above for classical languages.

Other circular concepts such as "reference," "satisfaction," and "exemplification" are governed by rules analogous to Aristotle's Rules for truth, and they receive a parallel treatment.[40] Vagueness can also be treated within a revision-theoretic framework.[41]

The scope of the Generality Requirement can be understood (and should be understood) to include concepts stipulatively defined using circular and interdependent definitions. Finite circular definitions, in particular, provide a good testing ground for a theory of truth, for their behavior is relatively clear and simple. Many otherwise attractive theories of truth (including some revision theories) do not lead to plausible theories of finite circular definitions. We believe that the theory we have offered here does so. We cite Theorem 3.3.2 as evidence.

[39] A further objection may be lodged against us. The very fact that we are introducing new conditionals, it may be said, is a problem for us. For we now need to provide an account of which ordinary uses of 'if' correspond to which conditionals—a task that is none too easy. *Response*: (i) If introducing new conditionals is a problem for us then it is a problem for all leading theories of truth, for they too find it necessary to introduce new conditionals. (ii) In ordinary uses, we do not distinguish—and, for the most part, it causes no trouble if we do not distinguish—the different conditionals. For if the constituent sentences are uniformly stable (i.e., stable in all revision sequences), the step-conditionals are equivalent to the material conditional. It is only in highly specialized contexts that there is a need to make distinctions, and it is the job of the theory of truth to spell out the distinctions and how they are to be deployed.

[40] See Gupta and Belnap [34, ch. 7], and Orilia [50].

[41] See Asmus [4].

3.6 A Comparison

We have offered above a theory of self-referential truth for a classical language equipped with additional logical resources—namely, the step biconditionals or, equivalently, $\Box$. Aczel and Feferman were perhaps the first logicians to explicitly set down and investigate theories of truth of this kind.[42] Aczel and Feferman studied, in particular, the type-free axiomatic system S($\equiv$), which they formulated in a classical language extended with a new biconditional ($\equiv$).[43] (They studied also some other systems closely related to S($\equiv$).) The system S($\equiv$) contains as axioms the Tarski biconditionals (and, more generally, axioms for satisfaction) with the biconditional connective read as their new equivalence ($\equiv$). Aczel and Feferman proved the consistency of S($\equiv$). We will not give here a detailed account of S($\equiv$), for which the reader should consult Aczel and Feferman [1] or, better, Feferman [22]. We note only some basic differences between reading the Tarski biconditionals in terms of the Aczel-Feferman biconditional and in terms of the step biconditional. For ease of comparison, we use $\equiv_{AF}$ for the Aczel-Feferman biconditional.[44]

(a) The biconditional $\equiv_{AF}$, unlike our step biconditional, expresses an equivalence relation. Consequently, in S($\equiv$), both of the following statements are provable for the Liar sentence, $\neg Tl$:

 (i) $Tl \equiv_{AF} (Tl \,\&\, \neg Tl)$, and
 (ii) $\neg Tl$.

 Since Tl is not provable, it follows that the theorems of S($\equiv$) are not closed under Truth Introduction: a sentence A can be a theorem while T(‘A’) fails to be a theorem. None of this holds for the step T-biconditionals. These biconditionals imply neither the Liar sentence, $\neg Tl$, nor the analogue of (i):

$$Tl \leftrightarrow (Tl \,\&\, \neg Tl).$$

 Furthermore, if A is provable from the step T-biconditionals then so also is $T(\text{‘}A\text{’})$.

(b) In light of the provability of (ii), it is plain that $A \equiv_{AF} B$ and B can be provable while A may fail to be provable. Not so for the step biconditional: if $A \leftrightarrow B$ and B are provable then A is also provable.[45]

(c) Aczel and Feferman’s new connective, $\equiv_{AF}$, expresses some sort of equivalence, but the exact character of this equivalence is unclear (as is noted by Feferman [22], Section 11). The step biconditional, in contrast, does not express

[42] Feferman points out that Heinrich Behmann informally explored as early as 1931 ideas that point toward these theories. See Feferman [22], section 14, which provides a brief history of type-free theories of truth and satisfaction.

[43] Aczel and Feferman’s notational conventions are the opposite of ours. Their symbol for material equivalence is $\leftrightarrow$, and their symbol for the new non-classical equivalence is $\equiv$.

[44] We wish to stress that Aczel and Feferman do not commit themselves to S($\equiv$). Indeed, they themselves have expressed reservations about it.

[45] That is, the inference rule “$A \leftrightarrow B$, B; therefore, A” is *admissible* in the sense explained above (Section 3.5.1(iii)). Note, however, that the rule does not hold unrestrictedly in hypothetical contexts.

an equivalence. In fact, the relation it expresses fails to be reflexive, symmetric, as well as transitive. Nonetheless, the step biconditional expresses a perfectly clear notion. Hence, while the meaning of the Tarski biconditionals is clear when they are interpreted using $\leftrightarrow$, their meaning is not clear when they are interpreted using $\equiv_{AF}$.

(d) S($\equiv$) seems to be a theory, set out in a classical language, of partial truth. (Indeed, Aczel and Feferman prove the consistency of a system very similar to S($\equiv$) via a Strong Kleene fixed-point construction.) Our theory is also formulated in a classical language. It is not, however, a theory of *partial* truth; it is a theory of *circular* truth. The step T-biconditionals have the character they do because they reflect the circularity of the concept of truth.

4 The Field Conditional

4.1 Exposition

Field's theory of truth combines fixed-point and revision-theoretic ideas in an interesting way. Field supplements the ground language with a truth predicate, T, and a conditional, which we will represent thus: $\rightarrow_F$. Field interprets the truth predicate via the least fixed-point of the Strong Kleene scheme, and the conditional using a revision construction, which we sketch below. The construction involves, as usual, iterative revisions of a hypothesis by a revision rule. Let us first make the idea of hypothesis precise; then we shall turn to Field's revision rule.[46]

Let M be, as before, a model of the ground language. Let $\mathbf{C}$ be the set of all formulas whose main connective is $\rightarrow_F$, and let V be the set of all assignments of values to variables relative to the domain of M. Then, *hypotheses h* for revision are certain functions from $(\mathbf{C} \times V)$ into the set of truth values $\{\mathbf{t}, \mathbf{f}, \mathbf{n}\}$.[47]

Let $M + h$ be an interpretation just like M except that it treats the conditional formulas as atomic and interprets them via the hypothesis h. We are now in a familiar three-valued setting and know, therefore, that there must be a Strong Kleene least fixed-point interpretation $F_{M,h}$ for truth relative to $M + h$. Let $M + h + F_{M,h}$ be the interpretation just like $M + h$ except that it assigns to the truth predicate T the interpretation $F_{M,h}$. We can now use the Strong Kleene rules to evaluate the entire language relative to $M + h + F_{M,h}$. (We interpret the conditional formulas, as before, using the hypothesis h.) Let $val_{M,h}$ be the function that assigns to each pair of formula and assignment its truth value in the resulting evaluation. Let $\leq$ be the following

[46] Our presentation of Field's ideas is based primarily on Field [23]. The study of Aristotle's Rules (and, more generally, comprehension principles) interpreted via new conditionals within non-classical logics has a long history. See Feferman [22] and Cantini [15] for references. See Dutilh Novaes [20] for a historical comparison. Recent contributors include Brady [10], Beall [7], Zardini [81], and Bacon [5].

[47] Field imposes a "local determination" condition on hypotheses: if a hypothesis assigns different truth values to a formula A relative to different assignments, then the assignments must differ over one or more variables free in A. See Field [23, p. 242].

linear ordering of the truth values:

$$\mathbf{f} \leq \mathbf{n} \leq \mathbf{t}.$$

Then, Field's revision rule, ϕ_M, is defined as follows. Let h be an arbitrary hypothesis, $(A \rightarrow_F B)$ an arbitrary conditional formula, and v an arbitrary assignment of values to variables. Then:

$$\phi_M(h)((A \rightarrow_F B), v) = \begin{cases} \mathbf{t} \text{ if } val_{M,h}(A, v) \leq val_{M,h}(B, v), \\ \mathbf{f} \text{ otherwise.} \end{cases}$$

Observe that the revised hypothesis assigns a classical truth value, **t** or **f**, to each element of its domain.

Let $h_{\mathbf{n}}$ be the constant hypothesis that assigns the value **n** to all pairs in $(\mathbf{C} \times V)$. Construct an On-long revision sequence, $\mathscr{S}$, that begins with $h_{\mathbf{n}}$, applies the rule ϕ_M at successor stages, and assigns to unstable elements of $(\mathbf{C} \times V)$ the value **n** at limit stages. This revision sequence must contain reflection stages α (see Section 3.1). Since the value of $\mathscr{S}$ at all reflection stages is bound to be the same, the following definition of h^* is legitimate: $h^* = \mathscr{S}_\alpha$. We shall call h^* the *reflection hypothesis* for M. Now, the *canonical model M^* over M* is this:

$$M + h^* + F_{M,h}.$$

This is the model Field uses to interpret the truth predicate and the conditional.[48]

Let us note that Field improves on Kripke's theory in one important respect. Unlike Kripke, he can recognize a sense of 'if' in which 'if A then A' and the Tarski biconditionals are logical truths. Set:

$$(A \leftrightarrow_F B) = (A \rightarrow_F B) \mathbin{\&} (B \rightarrow_F A).$$

Then, it can be verified that all sentences of the form

$$A \rightarrow_F A, \text{ and } T(\text{`}A\text{'}) \leftrightarrow_F A$$

are bound to be *valid* in this sense: they are true in the canonical model M^* over any ground model M.[49]

4.2 Some Observations

(i) The Field conditional differs semantically from our step conditionals in one important respect. Our step conditionals are *cross-stage* conditionals: in evaluating these conditionals, one considers the truth values of the antecedent and

[48] Field uses the construction just sketched for a consistency proof. He does not think that the construction spells out the real meaning of the truth predicate or of the conditional. Nonetheless, we focus on this construction because it is our best guide to Field's account of the conditional. Field calls the construction the "Official Theory" in Section 17.5 of his book.

[49] More precisely: a sentence B is *valid* by Field's semantics iff, for all ground models M, there is an assignment v of values to variables such that

$$val_{M,h^*}(B, v) = \mathbf{t},$$

where h^* is the reflection hypothesis for M. The notion of validity of arguments can be recovered, as before, from the notion just defined.

the consequent at different stages of revision. In contrast, the Field conditional is a *same-stage* conditional: in evaluating it, one considers the truth values of the antecedent and the consequent at the same stage.

(ii) We have seen that the Field conditional provides a reading of 'if' on which 'if A then A' is a logical law. Hence, a reading is available to Field on which argument (A3) is valid. However, Field can provide no reading of 'if' under which the law 'if A, then if B then $(A \mathbin{\&} B)$' is valid. Hence, the theory rules as invalid arguments closely related to (A3), for example:

(A3′) Suppose everything Fred says is true. Now suppose that everything Mary says is true. So, everything Fred says as well as everything Mary says is true. By conditional proof, if everything Fred says is true, then if everything Mary says is true then everything Fred says as well as everything Mary says is true.

Field's theory makes available four different readings of this argument: each of the two occurrences of 'if' in the conclusion can be interpreted either as $\supset$ or as $\rightarrow_F$. However, the argument is valid under none of the resulting readings.

(iii) According to Field's theory, argument (A10), which has the Truth-Teller as its premiss and 'snow is black' as its conclusion, is valid. It also deems valid analogs of (A10) constructed using the Field conditional, such as the following:

$$c = \top \rightarrow_F Tc, Tc;\ \text{therefore, } Snow\ is\ black.$$

This is not only counterintuitive; it puts an unbearable burden on the logic of truth. For it is mysterious how 'snow is black' could be deduced, using rules for truth, from the Truth-Teller. The problem could be overcome if Field were to base his construction on certain other three-valued fixed-points and certain other revision sequences. We wish to point out, however, that no choice of specific fixed-point and specific revision sequence is free of unwanted and troublesome validities. The only hope of overcoming the general difficulty is to quantify over a range of fixed points and a range of revision sequences.[50] If Field were to follow this course, he would need to modify his construction to bring it more in line with the theory we have offered above.[51]

(iv) Field does not provide a definite and precise account of validity even for the propositional fragment of the language. He tell us that

> 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 1 [i.e., **t**] in the given semantics. (Field [23, p. 277])

[50] Field partly recognizes this point in the course of his response to a related objection of Yablo [78].

[51] Certain unwanted validities are, however, forced by Field's general framework. For more on this point, see Standefer [67].

But this leaves wide open the logic of the Field conditional.[52,53]

4.3 Intersubstitutivity Principle

Our main difficulty with Field boils down to this: that the logic of the Field conditional is obscure; in contrast, the logic of our step conditionals is straightforward. Field has, however, an apparently powerful response: his theory validates the Intersubstitutivity Principle, but the theory we offer does not.

Intersubstitutivity Principle (IP): A and T(‘A’) are intersubstitutable in all extensional contexts. That is, if sentences B and C are exactly alike except that some extensional occurrences of T(‘A’) in one are replaced by A in the other, then B and C are inter-derivable; that is, C can be inferred from B, and B from C.[54]

This principle is bound to fail for us because our theory validates the classical law $(A \equiv A)$. Hence, our theory rules that (1) is valid:

(1) $l =$‘$\neg Tl$’ $\supset (T($‘$\neg Tl$’$)\equiv T($‘$\neg Tl$’$))$.

Now, if (IP) held, then (1) would be inter-derivable with (2):

(2) $l =$‘$\neg Tl$’ $\supset (T($‘$\neg Tl$’$) \equiv \neg Tl)$.

So, (2) would also be valid; this, however, is impossible. Hence, (IP) fails in our theory. This consequence, as Field sees it, is a fatal flaw. Any theory that fails to preserve (IP), Field thinks, ends up stripping the notion of truth of one of its essential functions. Field tells us:

> ‘true’ needs to serve as a device of infinite conjunction or disjunction (or more accurately, a device of quantification). ... in order for the notion of truth to serve its purposes, we need what I’ve been calling the *Intersubstitutivity Principle*. (Field [23, p. 210])

But what precisely is the connection between (IP) and the purposes served by ‘true’? Field’s thought here derives from an idea of W. V. Quine, who proposed that a function of ‘true’ is to enable us to generalize on sentence positions. Quine explains the point as follows in a famous passage in his *Philosophy of Logic*:

> We may affirm the single sentence by just uttering it, unaided by quotation or by the truth predicate; but if we want to affirm some infinite lot of sentences that we can demarcate only by talking about the sentences, then the truth predicate has its use. We need it to restore the effect of objective reference when for the

[52] For further discussion of Field’s logic, see Welch [75], McGee [48], Restall [59], and Priest [56].

[53] We wish to stress that the above critical observations are narrow in scope. Even if correct, they show at best that Field’s theory fails to meet the Descriptive Adequacy desideratum from Section 1. They do not by themselves provide a reason for rejecting Field’s theory. For, arguably, there are other important desiderata, and it could well be that the failure of Descriptive Adequacy is amply compensated by the success of Field’s theory in meeting these other desiderata. It is not our aim to offer here a full critique of Field’s theory.

[54] It is this principle that motivates the choice of the Strong Kleene scheme in Field’s construction.

> sake of some generalization we have resorted to semantic ascent. (Quine [57, p. 12])

Quine talks about affirmation here, but, as Field observes, the point is broader: 'true' serves the generalization function not only when truth attributions are affirmed outright, but also when they are embedded within compound sentences. Consider the following example from Field, in which a truth attribution occurs as the antecedent of a conditional:

(3) If everything that the Conyers report says is true then the 2004 election was stolen.

Suppose that the Conyers report says $A_1, \ldots, A_n$ and nothing more. Then, the generalization function of 'true' requires that (3) be inter-derivable with

(4) If A_1 & ... & A_n then the 2004 election was stolen.

Now, given the supposition, the rules governing quantification ensure that (3) is inter-derivable with

(5) If 'A_1' is true & ... & 'A_n' is true then the 2004 election was stolen.

So, to ensure the inter-derivability of (3) and (4), the logic of 'true' must render (4) and (5) inter-derivable. Field observes, correctly, that it will not do to confine the intersubstitutability of A and "'A' is true" to categorical contexts. The logical rules governing 'true' must allow us to move from (4) to (5), and back again. (IP) guarantees that the move is legitimate. Any theory that blocks the move ends up denying an essential function of 'true'.[55]

Two points should be noted about this argument. First, the argument makes plausible the idea that the generalization function of 'true' requires that (4) and (5) be inter-derivable. However, the argument does not establish the precise character of the inter-derivability that needs to obtain. If we run the argument in the simplest case, we arrive at the demand that an inter-derivability needs to obtain between A and "'A' is true." But, plainly, the character of this inter-derivability is not obvious. Indeed, our current inquiry, which has now occupied us for so long, is concerned principally to uncover its precise character.

Second, and this is the more important point, (IP) is not needed to ensure the inter-derivability of (4) and (5). The following weaker principle suffices.

Uniform Substitutivity Principle (UP): Let B be an arbitrary sentence in which all extensional occurrences of the truth predicate T are confined to contexts of the form T('A'). Let C be the formula that results when we replace *all* extensional occurrences of the form T('A') in B by A. (We shall call C the *uniform reduct* of B.) Then, B and C are inter-derivable.

(IP) allows for *mixed* substitutions of A for T('A'). That is, (IP) allows us to replace some occurrences of the form T('A') by A while leaving other occurrences

[55] Beall [7] and Cobreros *et al.* [19], among others, also accept the idea that (IP) is needed to sustain the generalization function of truth.

unchanged. (UP) is not so liberal: if we replace one occurrence, then the principle requires us to replace *all* extensional occurrences of the form T('A'). So, as we saw above, (IP) allows us to derive (2) from (1), and (1) from (2). (UP), on the other hand, allows no such derivations. It allows us to derive only the tautological

$$l = \text{`}\neg Tl\text{'} \supset (\neg Tl \equiv \neg Tl)$$

from (1).

This example brings out a vitally important difference between (IP) and (UP): in the presence of self-referential truth, (IP) forces us to deviate from classical logic; (UP), on the other hand, forces no such deviation.[56]

(IP) fails, we noted above, in the theory we have offered. (UP), we now wish to point out, is preserved. Indeed, we can now sharpen (UP): if C is the uniform reduct of B, then although

$$B \equiv C$$

fails to be valid, the sentence

$$B \leftrightarrow C$$

is valid. Indeed, the latter sentence is a logical consequence of the step T-biconditionals. In short, the generalization function of 'true' does not require the strong principle (IP); the weaker principle (UP) suffices. Hence, the generalization function creates no difficulty for our theory.[57]

The problems with the Field conditional are rooted, we think, in Field's misconceived adherence to (IP). It is this adherence that leads Field to burden his conditional with two separate tasks. The first task is to help overcome an expressive incompleteness in fixed-point theories—more specifically, to equip the language containing the truth predicate with a well-behaved conditional. The second task is to provide one or more readings of 'if' that validate Aristotle's Rules. These are two separate tasks, but they can seem to collapse into one if (IP) is accepted. (IP) implies that a reading of 'if' will validate Aristotle's rules so long as it validates Identity:

$$\text{if } A \text{ then } A.$$

The impression arises, therefore, that we can complete both tasks if we can add to the fixed-point language a well-behaved conditional, one that validates Identity. But this thought, rooted in (IP), is erroneous. The problem of interpreting the conditionals in Aristotle's Rules is quite separate from that of making sense of a well-behaved conditional in the presence of self-referential truth.[58] The first problem requires us

[56] For further discussion of, and a different perspective on, (IP) in the context of classical logic, see Cobreros *et al.* [19], and Ripley [60, 61].

[57] Note that restricted versions of (IP) also hold in our theory. So, for example, if A contains no occurrences of the truth predicate, then A and T('A') can be substituted for one another in all extensional contexts. For these kinds of substitutions, the distinction between (IP) and (UP) is of little significance and can be neglected. However, when self-referential truth is in play, the distinction between (IP) and (UP) is of vital importance. One needs to be very careful in extending ideas that hold for non-self-referential truth to truth in general.

[58] This point is confirmed by classical revision theories. These theories provide languages with a well-behaved conditional, but they do not provide conditionals suitable for interpreting Aristotle's Rules.

to discover readings of ‘if’ suitable for expressing certain laws governing circular concepts, laws such as Aristotle’s Rules. We showed that no ordinary conditional will do here (Section 2.2), and we went on to argue that two readings of ‘if’ are needed—our step-up and step-down conditionals. Neither of these conditionals, we have seen, validates Identity. The second problem requires us to make sense of truth in an expressively rich context, a context in which fixed-points do not exist. The solution of this problem, we have suggested, requires us to move beyond fixed-point ideas to revision theory. The Field conditional, burdened as it is with two disparate tasks, is unable to perform either one of them in a satisfactory way. The conditional is neither well-behaved nor does it provide a satisfactory reading of Aristotle’s Rules.

Acknowledgments We presented some of the ideas of this paper in talks given at Carnegie Mellon University, MIT, University of Barcelona, University of Illinois at Chicago, and the UConn Logic Seminar. We wish to thank our audiences for their questions and comments. We are especially grateful to the anonymous referees for their helpful suggestions.

References

1. Aczel, P., & Feferman, S. (1980). Consistency of the unrestricted abstraction principle using an intensional equivalence operator. In Hindley, R., & Seldin, J. (Eds.) *To H.B.Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism*, (pp. 67–98): Academic Press.
2. Antonelli, A. (1994). The complexity of revision. *Notre Dame Journal of Formal Logic*, *35*(1), 67–72.
3. Antonelli, A. (2002). The complexity of revision, revised. *Notre Dame Journal of Formal Logic*, *43*(2), 75–78.
4. Asmus, C.M. (2013). Vagueness and revision sequences. *Synthese*, *190*(6), 953–974.
5. Bacon, A. (2013). A new conditional for naive truth theory. *Notre Dame Journal of Formal Logic*, *54*(1), 87–104.
6. Barwise, J., & Etchemendy, J. (1987). *The Liar: An Essay on Truth and Circularity*: Oxford University Press.
7. Beall, J. (2009). *Spandrels of Truth*: Oxford University Press.
8. Belnap, N. (1982). Gupta’s rule of revision theory of truth. *Journal of Philosophical Logic*, *11*(1), 103–116.
9. Belnap Jr., N.D., & Grover, D.L. (1973). Quantifying in and out of quotes. In Leblanc, H. (Ed.) *Truth, Syntax and Modality, Studies in Logic and the Foundations of Mathematics* (Vol. 68, pp. 17–47): Elsevier.
10. Brady, R. (2006). *Universal logic*: CSLI Publications.
11. Bruni, R. (2013). Analytic calculi for circular concepts by finite revision. *Studia Logica*, *101*(5), 915–932.
12. Burge, T. (1979). Semantical paradox. *Journal of Philosophy*, *76*(4), 168–198.
13. Burgess, A.G., & Burgess, J.P. (2011). *Truth*: Princeton University Press.
14. Burgess, J.P. (1986). The truth is never simple. *Journal of Symbolic Logic*, *51*(3), 663–681.
15. Cantini, A. (2009). Paradoxes, self-reference and truth in the 20th century. In Gabbay, J.D. (Ed.) *Handbook of the History of Logic* (Vol. 5, pp. 875–1013): Elsevier.
16. Chapuis, A. (1996). Alternative revision theories of truth. *Journal of Philosophical Logic*, *25*(4), 399–423.
17. Chihara, C. (1979). The semantic paradoxes: A diagnostic investigation. *Philosophical Review*, *88*(4), 590–618.
18. Chihara, C.S. (1984). The semantic paradoxes: Some second thoughts. *Philosophical Studies*, *45*(2), 223–229.
19. Cobreros, P., Egre, P., van Rooij, R., & Ripley, D. (2013). Reaching transparent truth. *Mind*, *122*(488), 841–866.

20. Dutilh Novaes, C. (2008). A comparative taxonomy of medieval and modern approaches to liar sentences. *History and Philosophy of Logic*, *29*(3), 227–261.
21. Eklund, M. (2002). Inconsistent languages. *Philosophy and Phenomenological Research*, *64*(2), 251–275.
22. Feferman, S. (1984). Toward useful type-free theories. I. *Journal of Symbolic Logic*, *49*(1), 75–111.
23. Field, H. (2008). *Saving Truth from Paradox*: Oxford University Press.
24. Friedman, H., & Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, *33*, 1–21.
25. Gaifman, H. (1992). Pointers to truth. *Journal of Philosophy*, *89*(5), 223–261.
26. Gaifman, H. (2000). Pointers to propositions. In Chapuis, A., & Gupta, A. (Eds.) *Circularity, Definition, and Truth,* pp. 79 – 121: Indian Council of Philosophical Research.
27. Glanzberg, M. (2004). A contextual-hierarchical approach to truth and the liar paradox. *Journal of Philosophical Logic*, *33*(1), 27–88.
28. Gupta, A. (1982). Truth and paradox. *Journal of Philosophical Logic*, *11*(1), 1–60. A revised version, with a brief postscript, is reprinted in [43], pp. 175–235.
29. Gupta, A. (1988–89). Remarks on definitions and the concept of truth. *Proceedings of the Aristotelian Society*, *89*, 227–246. Reprinted in [33], pp. 73–94.
30. Gupta, A. (1993). A critique of deflationism. *Philosophical Topics*, *21*(2), 57–81. Reprinted with a postscript in [33], pp. 9–52.
31. Gupta, A. (1997). Definition and revision: A response to McGee and Martin. *Philosophical Issues*, *8*, 419–443. A revised version with a postscript is reprinted as “Definition and Revision” in [33], pp. 135–163.
32. Gupta, A. (2006). Finite circular definitions. In Bolander, T., Hendricks, V.F., & Andersen, S.A. (Eds.) *Self-Reference*, (pp. 79–93): CSLI Publications.
33. Gupta, A. (2011). *Truth, Meaning, Experience*: Oxford University Press.
34. Gupta, A., & Belnap, N. (1993). *The Revision Theory of Truth*: MIT Press.
35. Halbach, V. (2011). *Axiomatic Theories of Truth*: Cambridge University Press.
36. Herzberger, H.G. (1982). Notes on naive semantics. *Journal of Philosophical Logic*, *11*(1), 61–102. Reprinted in [43], pp. 133–174.
37. Horsten, L. (2011). *The Tarskian Turn: Deflationism and Axiomatic Truth*: MIT Press.
38. Kremer, P. (1993). The Gupta-Belnap systems $S^{\#}$ and S^{*} are not axiomatisable. *Notre Dame Journal of Formal Logic*, *34*(4), 583–596.
39. Kremer, P. (2010). How truth behaves when there’s no vicious reference. *Journal of Philosophical Logic*, *39*, 345–367. doi:10.1007/s10992-010-9136-4.
40. Kühnberger, K.U., Löwe, B., Möllerfeld, M., & Welch, P. (2005). Comparing inductive and circular definitions: Parameters, complexity and games. *Studia Logica*, *81*(1), 79–98.
41. Leitgeb, H. (2007). What theories of truth should be like (but cannot be). *Philosophy Compass*, *2*(2), 276–290.
42. Löwe, B., & Welch, P. (2001). Set-theoretic absoluteness and the revision theory of truth. *Studia Logica*, *68*(1), 21–41.
43. Martin, R.L. (1984). *Recent Essays on Truth and the Liar Paradox*: Oxford University Press.
44. Martinez, M. (2001). Some closure properties of finite definitions. *Studia Logica*, *68*(1), 43–68.
45. Martínez-Fernández, J. (2007). Maximal three-valued clones with the Gupta-Belnap fixed-point property. *Notre Dame Journal of Formal Logic*, *48*(4), 449–472.
46. Maudlin, T. (2004). *Truth and Paradox: Solving the Riddles*: Oxford University Press.
47. McGee, V. (1991). *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*: Hackett.
48. McGee, V. (2010). Field’s logic of truth. *Philosophical Studies*, *147*(3), 421–432.
49. Mendelson, E. (1997). *Introduction to Mathematical Logic, 4th edn*: Chapman and Hall.
50. Orilia, F. (2000). Property theory and the revision theory of definitions. *Journal of Symbolic Logic*, *65*(1), 212–246.
51. Parsons, C. (1974). The liar paradox. *Journal of Philosophical Logic*, *3*(4), 381–412.
52. Patterson, D. (2007). Understanding the liar. In Beall, J. (Ed.) *Revenge of the Liar: New Essays on the Paradox* (pp. 197–224): Oxford University Press.
53. Priest, G. (1979). The logic of paradox. *Journal of Philosophical Logic*, *8*(1), 219–241.
54. Priest, G. (1990). Boolean negation and all that. *Journal of Philosophical Logic*, *19*(2), 201–215.
55. Priest, G. (2006). *In Contradiction: A Study of the Transconsistent, 2nd edn*: Oxford University Press.
56. Priest, G. (2010). Hopes fade for saving truth. *Philosophy*, *85*(1), 109–140.

57. Quine, W.V.O. (1986). *Philosophy of Logic, 2nd edn*: Harvard University Press.
58. Restall, G. (2006). Minimalists about truth can (and should) be epistemicists, and it helps if they are revision theorists too. In Beall, J., & Armour-Garb, B. (Eds.) *Deflationism and Paradox* (pp. 97–106): Oxford University Press.
59. Restall, G. (2010). What are we to accept, and what are we to reject, while saving truth from paradox? *Philosophical Studies*, *147*(3), 433–443.
60. Ripley, D. (2013). Paradoxes and failures of cut. *Australasian Journal of Philosophy*, *91*(1), 139–164.
61. Ripley, D. (2013). Revising up: Strengthening classical logic in the face of paradox. *Philosophers' Imprint*, *13*(5).
62. Scharp, K. (2007). Aletheic vengeance. In Beall, J. (Ed.) *Revenge of the Liar: New Essays on the Paradox* (pp. 272–319):Oxford University Press.
63. Shaw, J.R. (2013). Truth, paradox, and ineffable propositions. *Philosophy and Phenomenological Research*, *86*(1), 64–104.
64. Simmons, K. (1993). *Universality and the Liar: An Essay on Truth and the Diagonal Argument*: Cambridge University Press.
65. Skyrms, B. (1984). Intensional aspects of semantical self-reference. In Martin, R.L. (Ed.) *Recent Essays on Truth and the Liar Paradox* (pp. 119–131): Oxford University Press.
66. Standefer, S. (2013). Truth, semantic closure, and conditionals. Ph.D. thesis, University of Pittsburgh.
67. Standefer, S. (2015). On artifacts and truth-preservation. *Australasian Journal of Logic*, *12*(3), 135–158. http://ojs.victoria.ac.nz/ajl/article/view/2045.
68. Standefer, S. (2015). Solovay-type theorems for circular definitions. *Review of Symbolic Logic*, *8*(03), 467–487.
69. Tarski, A. Der wahrheitsbegriff in den formalisierten sprachen. *Studia Philosophica*, *1*, 261–405. (1935/1983). An English translation by J. H. Woodger appears in [71] under the title "The concept of truth in formalized languages"; for full bibliographic information see the note on p. 152 of [71].
70. Tarski, A. (1944). The semantic conception of truth. *Philosophy and Phenomenological Research*, *4*(3), 341–376.
71. Tarski, A. (1983). *Logic, Semantics, Metamathematics*: Hackett.
72. Visser, A. (1984). Four-valued semantics and the liar. *Journal of Philosophical Logic*, *13*(2), 181–212.
73. Visser, A. (2004). Semantics and the liar paradox. In Gabbay, D., & Guethner, F. (Eds.) *Handbook of Philosophical Logic, vol. 11, 2nd edn.*, (pp. 149–240): Springer.
74. Welch, P.D. (2001). On Gupta-Belnap revision theories of truth, Kripkean fixed points, and the next stable set. *Bulletin of Symbolic Logic*, *7*(3), 345–360.
75. Welch, P.D. (2008). Ultimate truth vis-à-vis stable truth. *Review of Symbolic Logic*, *1*(01), 126–142.
76. Wintein, S. (2014). Alternative ways for truth to behave when there's no vicious reference. *Journal of Philosophical Logic*, *43*(4), 665–690. doi:10.1007/s10992-013-9285-3.
77. Woodruff, P.W. (1984). Paradox, truth and logic. *Journal of Philosophical Logic*, *13*(2), 213–232.
78. Yablo, S. (2003). New grounds for naive truth theory. In Beall, J. (Ed.) *Liars and Heaps: New Essays on Paradox*, (pp. 312–330): Oxford University Press.
79. Yaqūb, A.M. (1993). *The Liar Speaks the Truth: A Defense of the Revision Theory of Truth*: Oxford University Press.
80. Yaqūb, A.M. (2008). Two types of deflationism. *Synthese*, *165*(1), 77–106.
81. Zardini, E. (2011). Truth without contra(di)ction. *The Review of Symbolic Logic*, *4*(04), 498–535. doi:10.1017/S1755020311000177.