12

The Deflationist's Axioms for Truth

Volker Halbach and Leon Horsten

Aber es geht um die höhere Wahrheit, an die man glauben muß; und unsere Aufgabe ist es, diese Wahrheit in die Niederungen des Beweises herabzuziehen.

Hofrat Brunner to Ernst Stockinger
TV series *Stockinger* (episode *Stille Wasser*)

1. DEFLATIONISM

In this introduction we shall be very sketchy. We do not want to fatigue the reader by refuting in detail claims that have lost credibility a long time ago. For instance, we sketch only the arguments for the insufficiency of the T-sentences as axioms for truth.

We will state some claims that seem central to deflationism *as we understand it*. Naturally there will be philosophers who disagree with our conception of deflationism. We believe, however, that many will agree that the deflationist has to subscribe to these claims. They are weak in the sense that they describe more a methodology than a real philosophical doctrine. The claims are not intended to cover the deflationist position completely, and probably the deflationist will put forward much stronger claims.

In the first place, according to deflationism, a logico-mathematical notion of truth is central to the deflationist conception of truth. Thus whether a sentence or proposition is true does not depend on contingent facts such as our causal relations with the world. Consequently, for instance, a causal-historical notion of reference will not form the basis for a deflationist theory of truth. Rather truth behaves like a logical or mathematical expression: when it is combined with other logical and mathematical notions it forms sentences or propositions that obtain independently of any contingent facts in the world.[1] Presumably the deflationist will also need more 'substantial' notions of truth which are not logico-mathematical in this sense. But for the deflationist logico-mathematical truth is primary and the starting point from which other notions should be defined.

[1] See, e.g., Field [7], Halbach [13], Horsten [17].

Second, truth is axiomatized, that is, truth is conceived as a primitive and undefined notion. This approach does not exclude the possibility that truth turns out to be definable or reducible in another sense.[2] Deflationism, as we understand it, does not necessarily articulate a notion of truth that is 'thin' in the sense that it might easily be reduced away. Rather deflationists have tried to describe the purpose of truth that can only be achieved if truth is available. Or they say that truth would be dispensable if we could use infinite conjunctions or certain forms of quantification.[3] An axiomatization of truth also coheres with its status as a logico-mathematical notion.

According to the conception of deflationism outlined so far, *semantical* theories of truth like Kripke's fixed-point theory or the rule-of-revision theory[4] are not deflationist theories. These theories provide definitions of truth in set theory. Thus truth is no longer conceived as a primitive concept. Moreover, these semantical concepts of truth are dispensable, because they are definable. In this respect semantical theories are similar to 'substantial' theories of truth; most varieties of the correspondence, coherence and pragmatist theory are supposed to *define* truth in terms of states of affairs, correspondence, coherence, utility, etc.

The semantical conceptions of truth from Tarski to Kripke and the revision theory rely on the availability of a stronger metalanguage where truth is defined. Consequently these notions of truth are not universal in the sense of being notions of truth for the whole language (or at least its logico-mathematical part) that we are using.[5] Frequently they show how to add to an arithmetical language a truth predicate and how to expand the standard model of arithmetic to a model of the extended language with the truth predicate. Although this may be informative with respect to an analysis of the semantical paradoxes, it does not provide a notion of truth for the language (or, more precisely, for the theory) we are using. In general, studying toy languages and theories from a set-theoretic standpoint will not satisfy the deflationist because he is seeking a notion of truth for the language of the theory he is using. Concepts of truth for weak toy languages for which we can define truth by set-theoretic means are of little immediate use to the deflationist.

In sum, in a certain sense semantical approaches provide a more 'deflationary' picture of truth because they purport to define truth. So according to these semantical theories, truth is ultimately redundant because it is definable. However, definable

[2] The discussion has focused on a very special sense of reducibility. It is pretty obvious from Tarski's theorem on the undefinability of truth that truth will not be reducible in the sense that it is definable. Truth, however, might be reducible in another sense, and in proof theory many concepts of reduction have been discussed and applied. Several authors like Shapiro [27], Ketland [19] and Field [8] and Azzouni [1] have discussed whether the deflationist is committed to the conservativeness of his theory of truth. See Halbach [13].

[3] Field [8] and Azzouni [1] might be exceptions. They seem to believe that a truth theory ought to be conservative. See Halbach [13] for a discussion. We shall return to the discussion of conservativeness below.

[4] See Kripke [22] and Belnap and Gupta [2].

[5] Some deflationists disagree with this view. Soames [30], for instance, conceives Kripke's theory as a deflationist theory of truth.

notions of truth are not of primary interest to the deflationist because they are always just notions of truth for at best a part of our 'real' language.

Many axiomatic approaches are also formulated for toy languages and theories such that finally models for these axiomatic theories can be defined (in relatively weak theories). For instance, very often logicians add truth axioms to the language of Peano arithmetic. Although Peano arithmetic is very weak compared to our usual mathematical assumptions (set theory), these investigations are nevertheless relevant. If adding certain truth axioms to Peano arithmetic yields a consistent theory, adding axioms of the same kind to Zermelo-Fraenkel ought to produce a consistent theory as well. This is not guaranteed, however. For adding truth axioms usually increases the proof theoretic strength of a theory, and one cannot even obtain a proof of relative consistency. That is, in most cases we are not able to prove that set theory plus the truth axioms are consistent even if set theory itself is assumed to be consistent. In fact, the situation is sometimes even worse than that. Examples can be found of apparently innocent axioms concerning the notion of satisfaction which can consistently be added to Peano Arithmetic, but which result in an inconsistent system when they are added to the Zermelo-Fraenkel axioms.[6]

However, several considerations are independent from the chosen base language. If some truth axioms are inconsistent with Peano arithmetic, they will be inconsistent with set theory as well. Moreover, arithmetic is a convenient setting for the study of axiomatic truth theories because we have names of expressions at our disposal. Numerals (of codes of expressions) may serve the same purpose as quotational names in natural language. The language of set theory lacks such names (although they could easily be added). Therefore building truth theories on arithmetical theories is simply a convenient approach but arithmetic merely serves as an example here.

2. TARSKI'S THEORIES OF TRUTH

Many deflationists have advanced 'disquotational' axioms for truth. In particular, some have relied on some variety of the T- or disquotational sentences

$$\mathrm{T}\ulcorner A \urcorner \leftrightarrow A.$$

Here $\ulcorner A \urcorner$ is a name for the sentence A or its Gödel number. Horwich's theory is similar, but there $\ulcorner A \urcorner$ would be a name for the proposition that A.[7]

In order to avoid inconsistency one can restrict the instances to such sentences A that do not contain the truth predicate.[8] As Tarski has noted, the T-sentences are far too weak to prove any interesting generalization. In particular, if a base theory plus the T-sentences prove a generalization of the form $\forall x(A(x) \rightarrow \mathrm{T}x)$ then the base theory

[6] See Horsten [16].

[7] See Horwich [18].

[8] Some consistent restrictions of the T-scheme yield unwanted consequences. McGee [25] has shown that Horwich's idea of excluding only the 'bad' instances of the schema does yield neither a unique nor a satisfying theory of truth.

proves that there are at most n objects satisfying $A(x)$, that is, the base theory proves $\exists_n x A(x)$ for some fixed number n.[9]

The deductive weakness of the T-sentences is *our* motivation for embracing a compositional theory of truth. Other authors—like Davidson—had different reasons (like finite axiomatizability) for relying on Tarski's compositional axioms for truth. In the first axiom P is any n-place predicate symbol (except T) and val(x) represents the function that assigns to any closed term t its value.

$$\forall t_1 \dots t_n \, (\mathrm{T}\ulcorner P t_1 \dots t_n \urcorner \leftrightarrow P\mathrm{val}(t_1) \dots \mathrm{val}(t_n)) \tag{T1}$$

$$\forall A \in \mathcal{L} \, (\mathrm{T}\ulcorner \neg A \urcorner \leftrightarrow \neg \mathrm{T}\ulcorner A \urcorner) \tag{T2}$$

$$\forall A, B \in \mathcal{L} \, (\mathrm{T}\ulcorner A \wedge B \urcorner \leftrightarrow \mathrm{T}\ulcorner A \urcorner \wedge \mathrm{T}\ulcorner B \urcorner) \tag{T3}$$

$$\forall A(v) \in \mathcal{L} \, (\mathrm{T}\ulcorner \forall v A(v) \urcorner \leftrightarrow \forall t \mathrm{T}\ulcorner A(t) \urcorner) \tag{T4}$$

The first axiom says that for any string of closed terms $t_1, \dots, t_n$, P followed by this string of terms is true if and only if P applies to the values of these terms. In the case of arithmetic there is no difficulty in defining val( ). According to axiom T4 a universally quantified sentence $\forall v A(v)$ is true if and only if all its instances $A(t)$ for any closed term t are true. Thus T4 captures a substitutional understanding of the quantifier. In an arithmetical framework this approach is sound because there are closed terms for any object because there is a numeral for every number.

In the general case, where some objects might lack terms designating them, however, we would have to employ satisfaction instead of the unary truth predicate. What we say below would go through for satisfaction instead of truth as well. Employing truth instead of satisfaction makes our notation somewhat more perspicuous.

$\forall t_1$ expresses quantification over closed terms, while $\forall A \in \mathcal{L}$ expresses quantification over all sentences of the base language $\mathcal{L}$, i.e., the language without the truth predicate, and $\forall A(v) \in \mathcal{L}$ expresses quantification over all formulas of $\mathcal{L}$ with only-v free. Quantification in quotational contexts can be explained in different ways: we do not provide details here.

Axioms T2–T4 say that truth commutes with connectives and quantifiers. We do not provide an exact formulation because for our purposes a rough sketch ought to be sufficient.

If the axioms T1–T4 above are added to a theory like Peano arithmetic or some set theory it seems attractive not only to add these truth-theoretic axioms but to extend the axiom schemes to the new language with the truth predicate as well. In the literature there has been some discussion on these additional truth-theoretic axioms.[10] Most authors (including us) agree that the axiom schemes should be extended to the language with the truth predicate in the case of PA or ZF. The details are tricky and will not be dealt with in this paper.[11]

[9] This observation is basically Tarski's: see his [31]. The proof relies on the fact that a proof contains only finitely many instances of the T-schema. So the T-sentences prove only *finite* generalizations. This holds even if the allow the truth theory to contain all instances of the induction scheme (including those with T) assuming that, e.g., PA is our base theory.

[10] See Shapiro [27], Field [8] and Halbach [13].

[11] See Feferman [6] for more information on the role of axioms schemes in truth theories.

3. DESIDERATA FOR AXIOMATIC THEORIES OF TRUTH

Tarski's solution of the liar paradox was an undivided success in mathematical logic and opened the road to model theory. In philosophy of language it was only a partial success. Tarski himself did not think that his solution of the liar paradox applied to natural language.

The deflationist is confronted with the problem of having to come up with a less restrictive solution of the liar paradox. Of course the derivation of an inconsistency has to be blocked, but not in Tarski's coarse way. It is not in the scope of the present paper to discuss all proposals that have been made to this end. However, we shall discuss some of the main contenders.

Once the deflationist has decided in favor of an axiomatic approach to truth, he can attempt to articulate general features that any such theory *S* must possess in virtue of its logico-mathematical function in ordinary discourse. Here is the list of desiderata that we propose:

1. *S* must satisfy a requirement of naturalness and simplicity. It must contain as few ad hoc elements as possible.

2. *S* must explicate the compositional nature of truth. *S* must explain as fully as possible how the truth-value of a sentence is determined by the truth-values of its component parts.

3. *S* must prove as many (true) infinite conjunctions as possible. The chief reason for having a truth predicate is to express infinite conjunctions. But expressing infinite conjunctions is of little use if we are not able to prove many of them.[12]

4. The logic in which *S* reasons about the truth predicate must be the same as the logic under which the truth predicate is closed according to *S*. In quasi-technical terms: the *outer logic* of *S*, i.e., the set of the sentences provable in *S*, must equal the *inner logic* $\{A | S \vdash T\ulcorner A \urcorner\}$ of *S*.

5. It is desirable that classical logic is used. This applies not only to the outer logic, but—by the previous desideratum—to the inner logic as well.

These desiderata considerably overlap and were partly inspired by Michael Sheard's list of 'naive criteria' which he thinks theories of truth—both axiomatic and semantic—need to satisfy as much as possible.[13]

Our Desideratum 1 is a paraphrase of Sheard's third criterion, which he also calls the criterion of simplicity. Of course, it is not easy to put forward criteria for naturalness and simplicity. We think that the T-sentences are simple and that T1–T4 are simple and natural as well, though less simple than the T–sentences. T2–T4 express that the truth predicate commutes with connectives and quantifiers. Therefore T2–T4 describe a simple 'algebraic' property of truth.

Desideratum 2 implies that the truth predicate should commute with quantifiers and connectives. Therefore it is a generalization of Sheard's fifth principle which says

[12] See Halbach [11, 13].

[13] See Sheard [29].

that theories of truth should contain the Barcan principle $\forall A(v) \in \mathcal{L}(\mathrm{T}\ulcorner \forall v A(v)\urcorner \leftrightarrow \forall t \mathrm{T}\ulcorner A(t)\urcorner)$. Strictly speaking $A(t)$ is no subformula of $\forall v A(v)$, but the axiom would be compositional in the strict sense if a satisfaction predicate were employed. For the axiom would then say that $\forall_v A(v)$ is true if $A(v)$ is satisfied by all objects. Therefore we consider axioms as T4 above and C4 below as compositional. We shall have to discuss compositionality again below.

Desideratum 3 is not explicitly present in Sheard's list. It is to some extent entailed by Sheard's sixth principle, which he formulates tentatively, and which says that S should be arithmetically nonconservative over the arithmetical basis over which S is formulated.[14] Deflationists might object to adopting this desideratum if they believe that truth has to be a simple and innocent concept in every respect. We believe that deflationist truth is a tool for formulating and *proving* generalizations (or, if you like, infinite conjunctions).

Deflationist truth is as insubstantial as other logico-mathematical concepts like the concept of elementhood. It does not rely on any 'substantial' concepts like causality, correspondence, coherence or utility. But this does not force the deflationist in any way to maintain that a theory of truth ought to be deductively weak. This implies in particular that we do not expect a theory of truth to be conservative in any reasonable sense. Rather if the concept of truth is useful it will turn out to be proof-theoretically strong and irreducible. This is in accord with the deflationist doctrine that truth is an indispensable tool for making generalizations.

Desideratum 4 expresses that we want to capture truth for the language or theory we are using. In particular, we want to avoid asymmetries like axiomatizing in classical logic a notion of truth in partial logic. This desideratum implies Sheard's demand that provability should entail truth. For, if a sentence A is provable, then A is in the outer logic and there it ought to be contained in the inner logic as well; i.e., it should be (provably) true.

Desideratum 5 for a truth theory framed in classical logic is motivated by the deflationist conviction that truth is primarily a logico-mathematical notion. Historically, classical logic has emerged from a conscious attempt to explicate the form of mathematical reasoning. So if truth is indeed a logico-mathematical notion, then it ought to be governed by classical logic.

Desideratum 4 overrules the last desideratum. That is, if one opts for non-classical logic, e.g., partial logic, for the inner or the outer logic, then both, the inner and the outer logic, must be governed by partial logic.

4. BEYOND TARSKI

Once one drops Tarski's solution of the liar paradox and allows the truth predicate to apply to sentences containing the truth predicate, we see basically two ways to go: Either one sticks to classical logic or one adopts partial (or many-valued) logic or at least a partial conception of truth (which may be described in classical logic).

[14] Sheard remarks that under fairly general conditions, this feature of S is a consequence of S's containing the Barcan formula for T.

The main bulk of proposals consists in solutions based on non-classical logic. One can allow truth value gluts or gaps, or make even more severe incisions in classical logic. As pointed out above, the deflationist seeks an axiomatic approach, not a semantical approach. However, semantical theories can be used in order to motivate axiomatic approaches. A typical example is axiomatization of Kripke's approach in partial logic, e.g. Kremer [21]. Solutions based on nonclassical logic can be axiomatized in classical logic as well. In this case the truth theory itself is formulated in classical logic, but it describes a nonclassical concept of truth. Important examples of such systems are the variants of the *Kripke–Feferman* theory KF (see Feferman [6], Reinhardt [26] and Cantini [4]), which is motivated by Kripke's [22] fixed point theory with the strong Kleene scheme, and VF of Cantini [3], which is motivated by Kripke's [22] fixed point theory with the supervaluations scheme.

Typically these systems are inconsistent with either the "consistency" axiom

$$\forall A \in \mathcal{L}_{\mathrm{T}} \neg(\mathrm{T}\ulcorner \neg A\urcorner \wedge \mathrm{T}\ulcorner A\urcorner),$$

which excludes truth value gluts, or with the "completeness" axiom

$$\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{T}\ulcorner \neg A\urcorner \vee \mathrm{T}\ulcorner A\urcorner),$$

which excludes truth value gaps. Proof-theoretical investigations have shown that such theories are very good at proving infinite conjunctions. In fact, Feferman [6] has provided a proof-theoretic analysis of the weak and strong reflective closures of Peano arithmetic (variants of KF) by means of infinite conjunctions. Although Feferman's paper has gone mostly unnoticed by the deflationists, his paper certainly has advanced the understanding of the relation of infinite conjunctions and truth a great deal.

KF and its relatives are notorious for their asymmetry between internal and external logic: they describe a notion of partial truth in classical logic. One can prove in KF that the liar sentence L is not true, i.e., $\mathsf{KF} \vdash \neg\mathrm{T}\ulcorner L\urcorner$. Since KF is classical and $L \leftrightarrow \neg\mathrm{T}\ulcorner L\urcorner$ is provable in the basic theory of syntax, the liar sentence itself is a theorem of KF. That is, KF proves both the liar sentence *and* that the liar sentence is not true! Therefore

$$\{A \mid \mathsf{KF} \vdash A\} \not\subseteq \{A \mid \mathsf{KF} \vdash \mathrm{T}\ulcorner A\urcorner\},$$

i.e. the inner logic of KF does not coincide with its outer logic. According to KF, the extension of the truth predicate is closed only under *partial* logic. Indeed, KF was discovered by a conscious attempt to formalize Kripke's *semantical* inductive theory based on the strong Kleene-scheme. In sum, KF scores miserably on Desideratum 4.

One could (and should) formulate KF in partial logic outright.[15] That way, at least the inner logic of the resulting system would coincide with its outer logic,[16] and it would cohere better with the partial picture behind the theory.[17] But the fact remains that we have abandoned classical logic.

[15] We have tried to formulate KF in Strong Kleene logic in [14].

[16] The obvious way of formulating KF in partial logic results in a system which *differs* from the inner logic of KF. It is, according to [14] stronger than CT but weaker than KF. Cf. also Kremer [21] for a formulation of Kripke's theory in partial logic.

[17] Kripke [22] was ambiguous on precisely this point. Reinhardt [26] is more consistent here, and so is Soames [30].

KF is still compositional, if the concept of compositionality is liberalized. For KF features axioms like $\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{T}\ulcorner \mathrm{T}\ulcorner \dot{A}\urcorner\urcorner \leftrightarrow \mathrm{T}\ulcorner A\urcorner)$ saying that an atomic sentence T_n is true if and only if n is (the code of) a true sentence. Unlike the case of the quantifier axiom, using a satisfaction predicate will not render this axiom compositional in the strict sense. We tend to view A as a component of $\mathrm{T}\ulcorner A\urcorner$. In this sense KF is still compositional.[18]

Stronger systems like the above-mentioned system VF of Cantini [3], which formalizes the Kripkean construction based on the supervaluation scheme, have the same deficiencies as KF. The inner logic of VF also differs from its outer logic.[19] Moreover VF is not compositional and there is no known way to reformulate the VF axioms in a compositional manner. Proof-theoretic results suggest that this is not possible. In [12] it was conjectured that what compositionality for truth theories corresponds to predicativity for subsystems of analysis, i.e. with the aid of compositional truth theories we can motivate the arithmetical part of predicative analysis, but not more. If this thesis is sound, VF cannot be reformulated as a compositional system. For VF is impredicative. The irreducibly non-compositional feature of VF is the presence of a reflection axiom

$$\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{Bew}_P(\ulcorner A\urcorner) \rightarrow \mathrm{T}\ulcorner A\urcorner)$$

where P is some (weak arithmetical) theory formulated in the language $\mathcal{L}_{\mathrm{T}}$. This axiom implies that, e.g., $B \vee \neg B$ is true, even if B is some paradoxical sentence like the liar. Thus a disjunction may be true even if both disjuncts lack a truth value, but it also might lack a truth value. $B \vee B$ will neither be true nor false. VF even proves that neither the liar sentence L nor its negation $\neg L$ are true; nevertheless it proves that $L \vee \neg L$ is true. Thus the truth of a sentence does not only depend on the truth value of its components but also on the syntactical shape of the sentence. This is a clear violation of compositionality.

As we shall see, reflection axioms can be reduced in some cases to compositional axioms. But if Halbach's thesis on compositionality and predicativity holds, then in the case of VF this is not possible.

In favor of VF it is to be said that VF is one of the strongest natural truth theories known so far. Therefore VF scores best at Desideratum 3 among all truth systems.

5. CLASSICAL REFLECTIVE TRUTH

Because of the problems with KF and VF we propose a system describing a *classical* notion of truth. The system explicitly denies the existence of truth value gluts and gaps. Here we do not argue that classical concepts of truth are superior to partial notions. We simply presuppose that classicality is a desirable feature. At least we find reasoning in partial logic awkward and not natural and thus we do not want to use

[18] See Halbach [12] for further discussion.

[19] So we do not agree with Sheard that VF scores much higher than KF on Desideratum 4.

partial logic.[20] And if we use classical logic in order to axiomatize a concept of partial truth, then the awkward asymmetry of inner and outer logic arises as in the case of KF.

We emphasize, however, that we do not believe that there is a single best set of axioms for the deflationist. Tarski's theorem on the undefinability rules out the "best" system as inconsistent. All desiderata we have listed cannot be satisfied equally well at the same time. Therefore we cannot come to a final decision and settle for a single system.

An obvious way to generalize T1–T4 is simply to let the quantifiers not only range over formulas without the truth predicate but also over formulas with the truth predicate. This yields the following axioms:

$$\forall t_1 \dots t_n\, (\mathrm{T}\ulcorner Pt_1 \dots t_n\urcorner \leftrightarrow P\mathrm{val}(t_1)\dots \mathrm{val}(t_n)) \tag{T1}$$

$$\forall A \in \mathcal{L}_{\mathrm{T}}\, (\mathrm{T}\ulcorner \neg A\urcorner \leftrightarrow \neg \mathrm{T}\ulcorner A\urcorner) \tag{C2}$$

$$\forall A, B \in \mathcal{L}_{\mathrm{T}}\, (\mathrm{T}\ulcorner A \wedge B\urcorner \leftrightarrow \mathrm{T}\ulcorner A\urcorner \wedge \mathrm{T}\ulcorner B\urcorner) \tag{C3}$$

$$\forall A(v) \in \mathcal{L}_{\mathrm{T}}\, (\mathrm{T}\ulcorner \forall v A(v)\urcorner \leftrightarrow \forall t \mathrm{T}\ulcorner A(t)\urcorner) \tag{C4}$$

We call the base theory plus these axioms CT_0 (for "classical truth").

We cannot allow P to be truth predicate in the first axiom, because otherwise the system would be inconsistent.[21]

The deflationist surely wants his theory to be sound, that is, he wants all theorems of his theory to be true. But of course Gödel's second incompleteness theorem shows that a truth theory can hardly prove its own soundness. The soundness for a truth theory S is expressed by:

$$\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{Bew}_s(\ulcorner A\urcorner) \rightarrow \mathrm{T}\ulcorner A\urcorner) \tag{GRFN}$$

This sentence is called the *global reflection axiom* for S.[22] If the truth theory proves the unproblematic T-sentence $\mathrm{T}\ulcorner \bot\urcorner \leftrightarrow \bot$, then GRFN implies also the consistency statement $\neg\mathrm{Bew}_s(\ulcorner \bot\urcorner)$ for S.

However we are free to add the reflection axiom GRFN to the truth theory S in order to obtain a new theory S_1. The result of iteratively adding reflection axioms to a theory has been investigated thoroughly.[23] However, usually soundness is not directly expressed because a truth predicate is not available. Therefore proof theorists confine themselves to the uniform reflection scheme

$$\mathrm{Bew}_s(\ulcorner A(\dot{\vec{x}})\urcorner) \rightarrow A(\vec{x}), \tag{RFN}$$

[20] We could invoke our Desideratum 1. But that comes down to the claim that we reject partial logic because it is not natural according to our taste. Others who are more used to partial logic will disagree. In the context of mathematics at least, the adoption of partial logic means an essential departure from our usual methods of reasoning that one should think twice before giving up classical logic.

[21] See Halbach [10].

[22] See Kreisel and Lévy [20].

[23] See Feferman [5].

which is supposed to express that all theorems of S are true. (The dot above x indicates that this variable is bound from outside in the usual way by formally substituting numerals for the variable x). In the present set-up we can avail ourselves of a truth predicate and we can express soundness by GRFN instead of its surrogate RFN.

The theory CT_0 has been defined above: it contains the base theory and the axioms postulating that the truth predicate commutes with quantifiers and connectives. We define recursively new systems which are obtained by adding the global reflection axiom for the respective system in the style of Turing's and Feferman's progressions.[24] CT_{n+1} is the system CT_n plus the global reflection axiom for CT_n:

$$\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{Bew}_{\mathrm{CT}_n}(\ulcorner A \urcorner) \rightarrow \mathrm{T}\ulcorner A \urcorner) \tag{CTR}$$

Since we are dealing with finite progression only and CT_{n+1} extends CT_n only by a single axiom, there is always a canonical provability predicate available and we do not have to deal with the problems of the intensionality of progressions.

CT is defined as the union of all theories $\mathrm{CT}_n (n \in \omega)$.

For the sake of definiteness we consider Peano arithmetic as the base theory and state a few facts about CT built over Peano arithmetic. The observations carry over to many other base theories.

First we compare CT to theories that have been discussed in the literature. To this end we remark that adding all the reflection principles to CT_0 is equivalent to adding the rule

$$\frac{A}{\mathrm{T}\ulcorner A \urcorner} \tag{NEC}$$

to CT_0, where A is any sentence in $\mathcal{L}_{\mathrm{T}}$.

The rule NEC can be derived in CT as follows. If A is derivable in CT_0, then—by the Σ_1-completeness of PA$-\mathrm{Bew}_{\mathrm{CT}_0}(\ulcorner A \urcorner)$ is derivable in PA, and by the global reflection principle for CT_0 we have $\mathrm{T}\ulcorner A \urcorner$. Iterated applications of NEC can the dealt with in a similar way.

Deriving the global reflection principles from NEC is slightly harder; we shall only sketch the proof. CT_0 proves that all induction axioms in $\mathcal{L}_{\mathrm{T}}$ are true because

$$\forall A(v) \in \mathcal{L}_{\mathrm{T}}(\mathrm{T}\ulcorner A(0) \urcorner \wedge \forall x(\mathrm{T}\ulcorner A(\dot{x}) \urcorner \rightarrow \mathrm{T}\ulcorner A(\dot{x}+1) \urcorner) \rightarrow \forall x \mathrm{T}\ulcorner A(\dot{x}) \urcorner)$$

is an instance of the induction scheme and implies that all induction axioms are true. Beyond the induction axioms CT_0 has only finitely many axioms. They can be proved to be true by applying NEC to any of these axioms. Since truth is provably closed under logic, we therefore can prove in $\mathrm{CT}_0 + \mathrm{NEC}$ that all theorems of CT_0 are true. For the induction step we have to show $\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{Bew}_{\mathrm{CT}_n}(\ulcorner A \urcorner) \rightarrow \mathrm{T}\ulcorner \mathrm{T}\ulcorner A \urcorner \urcorner)$ because this implies $\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{Bew}_{\mathrm{CT}_{n+1}}(\ulcorner A \urcorner) \rightarrow \mathrm{T}\ulcorner A \urcorner)$. By induction hypothesis we have already $\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{Bew}_{\mathrm{CT}_n}(\ulcorner A \urcorner) \rightarrow \mathrm{T}\ulcorner A \urcorner)$ and by one application of NEC and the CT_0-axioms $\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{T}\ulcorner \mathrm{Bew}_{\mathrm{CT}_n}(\ulcorner A \urcorner) \urcorner \rightarrow \mathrm{T}\ulcorner \mathrm{T}\ulcorner A \urcorner \urcorner)$. Since CT_0 proves all T-sen-

[24] See Turing [33] and Feferman [5].

tences for arithmetical instances we have $\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{Bew}_{\mathrm{CT}_n}(\ulcorner A \urcorner) \to \mathrm{T}\ulcorner \mathrm{T}\ulcorner A \urcorner\urcorner)$ as desired.

Thus the global reflection axioms and NEC are interderivable. This implies that CT is equivalent to the system FS without the converse rule of NEC.[25] That is, if we add

$$\frac{\mathrm{T}\ulcorner A \urcorner}{A}$$

to CT, then we obtain FS. Since it is still an unsolved problem whether this rule can be dropped from FS without any loss, for all we know CT and FS could be equivalent.[26] In fact, it would be nice if the system CT should turn out to be symmetrical, i.e., if the inner and outer logic of FS coincide. If it does not, we would recommend to add the above rule.

Friedman and Sheard have proved the consistency of FS; therefore CT is consistent as well. Models for CT may be obtained by revision semantics in the style of Herzberger.[27]

The iterated global reflection axioms of CT invite an obvious question: Why not iterate the reflection principle into the transfinite in the style of Feferman's transfinite progressions?[28] There is a very good reason for not doing this. The reflection axiom

$$\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{Bew}_{\mathrm{CT}}(\ulcorner A \urcorner) \to \mathrm{T}\ulcorner A \urcorner)$$

for CT is inconsistent with CT.[29] This follows from a result by McGee.[30] He showed that a subsystem of CT is ω-inconsistent. McGee's argument can easily be carried out inside CT, that is, CT proves $\mathrm{Bew}_{\mathrm{CT}}(\ulcorner \exists x \neg A(x) \urcorner$ and $\forall x(\mathrm{Bew}_{\mathrm{CT}}(\ulcorner A(\dot{x}) \urcorner)$ for some formula $A(x)$ of $\mathcal{L}_{\mathrm{T}}$. Together with the global reflection principle for CT the latter implies in CT_0 the truth of $\forall x A(x)$. Thus there is a very good reason for not iterating global reflection into the transfinite.

Sometimes it is thought that the ω-inconsistency of CT and similar systems shows that they are not attractive as axiomatizations of truth. We do not share this view.

CT can be shown to be arithmetically sound, that is, CT does not prove any false arithmetical sentence. So the ω-inconsistency concerns only the part of CT that deals with the truth predicate.

Why would one reject an ω-inconsistent truth theory? Why would one reject CT in particular? One cannot put one's hopes on any particular sentence. For the sentences involved in the ω-inconsistency are circular and truth theories generally disagree on such sentences. So one cannot reject CT because it proves a false sentence. Perhaps one

[25] The system FS ("Friedman–Sheard") was studied under a different name and proved consistent by Friedman and Sheard [9]. See also Halbach [10].

[26] Results by Halbach [10] show that this rule does not contribute to the proof-theoretical strength of FS and that only very special sentences could require this rule for their proof. Further results by Sheard [28], however, showed that this rule is not as weak as it may appear.

[27] See Herzberger [15].

[28] See Feferman [5].

[29] Indeed the uniform reflection principle for CT is already inconsistent with CT.

[30] See [24].

might argue that CT does not have a nice model and that any attractive theory of truth must possess a nice model. To us this line of argument appears questionable. We accept set theory although we cannot prove that there is any nice model for set theory. Because of Gödel's second incompleteness theorem we cannot even prove that set theory has any model. However, there is an important difference between the case of set theory and CT. For set theory does not rule out that there is a nice model of set theory. For instance, set theory does not refute the existence of an inaccessible cardinal number. The truth theory CT, however, is ω-inconsistent, which refutes the existence of a ω-model of CT. Thus one can see from within CT that any model of CT must be nonstandard and that CT is inconsistent with the uniform reflection principle for CT.

Nevertheless we do not think that this makes CT unacceptable. For the semantics of CT are not as weird as it might appear. In fact CT has a very natural semantics. For any subsystem of CT with a finite number of global reflection axioms possesses a nice standard model. Since we can use in any proof only finitely many reflection axioms, at any step of our reasoning we have a nice model. The model is provided by rule-of-revision semantics.[31] A model for CT_n can be obtained by $n + 1$ applications of the revision operator to the standard model of arithmetic. We sketch the procedure: Expand the standard model of arithmetic to the language $\mathcal{L}_{\mathrm{T}}$ by assigning an arbitrary extension S_0 to the truth predicate. The model has the form $(\mathbb{N}, S_0)$, where $\mathbb{N}$ is the standard model of arithmetic and S_0 is the chosen extension of the truth predicate. Given S_n define S_{n+1} in the following way: $S_{n+1} = \{A \in \mathcal{L}_{\mathrm{T}} | (\mathbb{N}, S_n) \models A\}$ (here we identify sentences with their codes), that is, we use the set of all sentences truth in the model $(\mathbb{N}, S_n)$ as the new extension of the truth predicate. This way we obtain models for the theories CT_k.

The problems of revision semantics at limit levels are well known. The problem arises because the set S_ω of sentences that stay in the extension of the truth predicate from some level on is ω-inconsistent.[32] Therefore there is no ω-model for S_ω and the only option that remains is taking S_ω as the new extension of the truth predicate.[33] The ω-inconsistency of CT just reflects the fact that there is no nice limit model at level ω in rule-of-revision semantics. We believe that the axiomatic approach proves superior to the semantical approach based on the standard model. For on the semantical side there is no nice limit model at level ω, while one can simply take the union of all systems CT_n. The consistency of every system CT_n ensures the consistency of the entire system CT.

Therefore at any step in a proof in CT we may affirm that anything we have claimed so far is true: we do affirm it by a global reflection principle. We cannot, however, reflect on this and conclude that all affirmations of the soundness must be sound as well. This would be an additional reflection step that would take us to the global reflection principle for CT itself, which is inconsistent with CT.

We shall now look again at our desiderata and check to what extent they are met by CT. We hope that we have succeeded in describing the axioms of CT as simple and

[31] See Belnap and Gupta [2].

[32] This phenomenon is discussed by Belnap and Gupta [2].

[33] This is basically Herzberger's [15] limit rule.

natural. In particular, the axioms of CT naturally extend the unquestionable axioms T1–T4 to axioms for self-applicative truth.

CT, as it stands, is not compositional. For the reflection axioms CTR are not compositional. But CT can be reformulated with compositional axioms and rules. For the rule NEC surely is compositional if the KF axiom $\forall A \in \mathcal{L}_{\mathrm{T}}(\mathrm{T}\ulcorner \mathrm{T}\ulcorner \dot{A} \urcorner\urcorner \leftrightarrow \mathrm{T}\ulcorner A \urcorner)$ was. KF rejects the compositional axiom C2 for negation, but C2 forms part of the axioms of CT. Thus with respect to the compositionality Desideratum 2 CT scores higher than KF.

The above model-theoretic considerations can be transformed into a proof of the fact that the proof-theoretic strength of CT is the same as of ramified analysis for all finite levels.[34] By comparison KF reaches the strength of ramified analysis up to ε_0. Therefore KF (and a fortiori VF) proves more generalizations than CT.

The real strength of CT lies in Desiderata 4 and 5. CT describes in classical logic a classical notion of truth. We have mentioned above that it is unknown whether

$$\frac{\mathrm{T}\ulcorner A \urcorner}{A}$$

is a derived rule of CT. If it were, the inner and outer logic of CT would be identical. This makes a proof of the hunch that this rule is derivable even more desirable. At any rate we can simply add this rule to CT in order to force the identity of inner and outer logic.[35] In contrast, adding NEC to KF in order to force the inner and outer logic of KF to be identical results in an inconsistent system.

6. CONCLUSION

When certain tenets of deflationism are accepted, one is driven to the axiomatic approach to the notion of truth. And then the question arises how the details of what is from a deflationist perspective the most attractive theory of truth would look like.

This question has hitherto not been given as much attention as it deserves. Deflationists have mostly adopted either Tarski's compositional theory of truth or, more frequently, Tarski's disquotational theory of truth. While this latter theory is simply deductively too weak, there presently exist also serious rivals to Tarski's compositional theory of truth, which are obtained by extending Tarski's compositional theory. In these theories one can prove that certain sentences containing the truth predicate are true. These systems do not obey Tarski's strict distinction between object- and metalanguage. They score much higher on Desideratum 3 than Tarski's compositional theory: they prove more generalizations.

However, one has to pay a price for the gain in expressive and deductive power: Many of these systems like KF and VF no longer describe a classical notion of truth; instead they describe a notion of truth with truth value gaps or gluts.

[34] See Halbach [10].

[35] But if CT + CONEC $\neq$ CT, then the question whether CT + CONEC can be analysed in terms of reflection principles needs to be investigated.

We have focused on a rival theory—CT—that sticks to a classical conception of truth thereby excluding truth value gaps and gluts. This system in many respects looks like a very natural strengthening of Tarski's compositional theory. Nevertheless, as soon as McGee discovered that this theory is ω-inconsistent, it was put aside.

In this chapter, we have argued that this was a hasty judgement. As a theory of truth, CT has much more to be said for it than is commonly appreciated. The effects of the ω-inconsistency are limited to the sphere of the diagonal sentences involving T, where our intuitions about the notion of truth are pretty much of no use anyway. And outside the sphere of these sentences, CT gives us only patently correct results. Moreover, it gives us many of them: CT is proof-theoretically significantly stronger than Tarski's compositional theory. For the moment the system CT (or FS if they are different) seems to be the deflationist's best bet. It is the most successful axiomatic theory of truth that is currently on the table.

REFERENCES

[1] Jody Azzouni. Comments on Shapiro. *Journal of Philosophy*, 96: 541–4, 1999.

[2] Nuel Belnap and Anil Gupta. *The Revision Theory of Truth*. MIT Press, Cambridge, 1993.

[3] Andrea Cantini. A theory of formal truth arithmetically equivalent to ID_1. *Journal of Symbolic Logic*, 55: 244–59, 1990.

[4] ——. *Logical Frameworks for Truth and Abstraction. An Axiomatic Study*, vol. 135 of *Studies in Logic and the Foundations of Mathematics*. Elsevier, Amsterdam, 1996.

[5] Solomon Feferman. Transfinite recursive progressions of axiomatic theories. *Journal of Symbolic Logic*, 27: 259–316, 1962.

[6] ——. Reflecting on incompleteness. *Journal of Symbolic Logic*, 56: 1–49, 1991.

[7] Hartry Field. Deflationist views of meaning and content. *Mind*, 103: 247–85, 1994.

[8] ——. Deflating the conservativeness argument. *Journal of Philosophy*, 96: 533–40, 1999.

[9] Harvey Friedman and Michael Sheard. An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, 33: 1–21, 1987.

[10] Volker Halbach. A system of complete and consistent truth. *Notre Dame Journal of Formal Logic*, 35: 311–27, 1994.

[11] ——. Disquotationalism and infinite conjunctions. *Mind*, 108: 1–22, 1999.

[12] ——. Truth and reduction. *Erkenntnis*, 53: 97–126, 2000.

[13] ——. How innocent is deflationism? *Synthese*, 126: 167–94, 2001.

[14] ——and Leon Horsten. Axiomaticing Kripke's theory of truth. Forthcoming.

[15] Hans G. Herzberger. Notes on naive semantics. *Journal of Philosophical Logic*, 11: 61–102, 1982.

[16] Leon Horsten. Concerning the notion of satisfaction. *Logique et Analyse*, forthcoming.

[17] ——. The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In P. Cortois, ed., *The Many Problems of Realism*, vol. 3 of *Studies in the General Philosophy of Science*, pp. 173–87. Tilburg University Press, Tilburg, 1995.

[18] Paul Horwich. *Truth*. Oxford University Press, Oxford, second edn, 1998, first edn 1990.

[19] Jeffrey Ketland. Deflationism and Tarski's paradise. *Mind*, 108: 69–94, 1999.

[20] Georg Kreisel and Azriel Lévy. Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 14: 97–142, 1968.

[21] Michael Kremer. Kripke and the logic of truth. *Journal of Philosophical Logic*, 17: 225–78, 1988.

[22] Saul Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72: 690–712, 1975. Reprinted in [23].

[23] Robert L. Martin, ed. *Recent Essays on Truth and the Liar Paradox*. Clarendon Press and Oxford University Press, Oxford and New York, 1984.

[24] Vann McGee. How truthlike can a predicate be? A negative result. *Journal of Philosophical Logic*, 14: 399–410, 1985.

[25] ——. Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic*, 21: 235–41, 1992.

[26] William Reinhardt. Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic*, 15: 219–51, 1986.

[27] Stewart Shapiro. Proof and truth: Through thick and thin. *Journal of Philosophy*, 95: 493–521, 1998.

[28] Michael Sheard. Weak and strong theories of truth. *Studia Logica*, 68: 89–101, 2001.

[29] ——. Truth, probability, and naive criteria. In Volker Halbach and Leon Horsten, eds, *Principles of Truth*. Dr. Hänsel-Hohenhausen, Frankfurt-am-Main, 2002.

[30] Scott Soames. *Understanding Truth*. Oxford University Press, New York and Oxford, 1999.

[31] Alfred Tarski. Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica*, 1: 261–405, 1935. Reprinted as "The Concept of Truth in Formalized Languages" in [32], pp. 152–278; page references are given for the translation.

[32] ——. The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics*, pp. 152–278. Clarendon Press, Oxford, 1956.

[33] Alan Turing. Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, 45: 161–228, 1939.