Axioms for determinateness and truth

Solomon Feferman

Abstract A new formal theory **DT** of truth extending **PA** is introduced, whose language is that of PA together with one new unary predicate symbol T(x), for truth applied to Gödel numbers of suitable sentences in the extended language. Falsity of x, F(x), is defined as truth of the negation of x; then the formula D(x) expressing that x is the number of a determinate meaningful sentence is defined as the disjunction of T(x) and F(x). The axioms of **DT** are those of **PA** extended by (I) full induction, (II) strong compositionality axioms for D, and (III) the recursive defining axioms for T relative to D. By (II) is meant that a sentence satisfies D if and only if all its parts satisfy D; this holds in a slightly modified form for conditional sentences. The main result is that **DT** has a standard model. As an improvement over earlier systems developed by the author, **DT** meets a number of leading criteria for formal theories of truth that have been proposed in the recent literature, and comes closer to realizing the informal view that the domain of the truth predicate consists exactly of the determinate meaningful sentences.

1. **Background** Much work has been devoted since the 1970s to breaking the binds of Tarskian hierarchies for truth. There are two basic approaches to non-hierarchical theories of truth, semantical and axiomatic. Work as of the early 1990s on both of these was usefully surveyed by Michael Sheard (1994), and as far as I know there has been no comparable canvass of the two approaches that would bring us up to date on the progress that has been made in the intervening years. Much information on that, however, can be gleaned from the piece by Hannes Leitgeb (2007) on criteria for theories of truth, of which more in the final section below. And on the axiomatic side we have the very useful entry by Volker Halbach (2007) in the online Stanford Encyclopedia of Philosophy.

This paper is intended as a further contribution to the axiomatic approach that in various respects is an improvement on an earlier effort I made in that direction. Some

background is needed to explain in just what respects that is and to motivate the present development. There are only two papers that I've written directly involving axiomatic theories of truth, namely "Toward useful type-free theories. I" (1984) and "Reflecting on incompleteness" (1991), referred to in the following respectively as (F 84) and (F 91). Both of these primarily concern non-hierarchical theories, but their purposes were quite different, as I will explain in a moment. However, there were aspects of each with which I was dissatisfied, and so I gave a good deal of thought off and on since their publication to obtaining more satisfactory theories. I arrived at the one presented here in the fall of 2005, prompted by an invitation to deliver the Tarski lectures at UC Berkeley in April 2006. I had decided that for the first lecture, nothing would be more fitting than to take truth, both as dealt with by Tarski and in certain of its non-hierarchical versions, as the central topic.[1] In the event, what I ended up with is—in my opinion—a formally more elegant system than what had been presented in the 1984 paper and one that is more satisfying philosophically.

Both (F 84) and (F 91) introduced axiomatic systems formulated in classical two-valued logic that were based on the well-known Kripke construction (1975) of a three-valued model for a non-hierarchical truth predicate. Of these, only the one in (F 91) closely axiomatizes that construction itself, and for that reason it has (with minor modifications) come to be referred to as the **KF** system (or Kripke-Feferman axioms). The Halbach entry (2007), sec. 4.3, gives some of the history surrounding that. Basically, I had presented it in two lectures for the Association of Symbolic Logic, the first in 1979 and the second in 1983, and circulated drafts of the material at that time. It was through those presentations that William Reinhardt (1985, 1986), Andrea Cantini (1989) and Vann McGee (1991) came to know and write about **KF** prior to the appearance of (F 91).

The aim of (F 91) was to put **KF** in service of a fairly general notion of *reflective closure* of an open-ended schematic theory—of what notions and principles one ought to accept if one accepts the basic notions and principles of the theory. This was explained in terms of reflection principles derived most naturally from iterating a non-hierarchical

[1] The lecture itself was entitled "Truth unbound"; it has not been published. The elaboration of the new ideas as given here was presented to the Workshop on Mathematical Methods in Philosophy held at the Banff International Research Station (18-23 February 2007) under the title, "A nicer formal theory of non-hierarchical truth". I am indebted to Kentaro Fujimoto, Hannes Leitgeb and Thomas Strahm for their useful comments on earlier drafts of the present article.

truth predicate. In particular, it was shown in that paper that the reflective closure (in its widest sense) of a schematic version of the Peano axioms for 0, successor and predecessor is of the same strength as predicative analysis. But I always thought that the **KF** axioms were a bit artificial for that purpose. Subsequently, in my paper (1996) I obtained a new notion for the same general purpose that I called the *unfolding* of an open-ended schematic theory. This does not require use of a truth predicate, is potentially more widely applicable than reflective closure and is, to my mind, more natural. In particular, Thomas Strahm and I have shown in our joint paper (2000) that the full unfolding of the system of non-finitist arithmetic (the basic schematic Peano axioms) is again of the same strength as predicative analysis. Strahm and I now have work in progress on the unfolding of finitist arithmetic. I mention all this in order to explain in what way (F 91) and its relatively familiar **KF** axioms are less relevant to the following.

So now let's turn back to the (F 84) paper. That presented some classical axiomatic theories that simultaneously generalized type-free theories of truth and type-free theories of membership. As its title indicates, the purpose was pragmatic. Namely, on the set-theoretical side, there are natural type-free statements that one would like to make about structures whose members are structures and that are not directly accounted for in standard theories of sets, or sets and classes. Category theory abounds in such examples, but one needn't go to anything so fancy to illustrate the idea. Consider, for example, the partially ordered structures; the collection of all these forms a partially ordered structure under the substructure relation, and so should be considered a member of itself. For another example, the structure consisting of all semi-groups forms a semi-group under Cartesian product, up to isomorphism. The planned Part II of "Toward useful type-free theories" was to contain such applications of the axiomatic systems on the set-theoretical side; however, that was never published since the applications did not work out as well as anticipated. In the meantime, I have explored alternative approaches for the same ends (see Feferman (2004), (2006)).

Though the axiomatic systems in (F 84) covered both theories of truth and membership, the possible applications on the former side would be quite different, and it muddied the picture to treat them together. Since these systems deserve reconsideration, another reason for dealing with truth first, as is done here, is that it is in certain respects

formally simpler. Let me begin with the informal guiding ideas. As laid out in (F 84) p. 78, if one is to have a consistent theory of truth, there are three possible routes that may be taken in the face of the paradoxes, namely by restriction of (1°) language, (2°) logic, or (3°) basic principles. Examples of (1°) are hierarchical theories of truth, such as Tarski's; since the aim here is for a non-hierarchical theory, that route is not taken. Examples of (2°) are systems of logic based on three values; in (F 84) p. 95 I argued that "nothing like sustained ordinary reasoning can be carried on" in the familiar such systems that are on offer. In the meantime, a different kind of restriction has been proposed via the use of so-called *paraconsistent* systems, i.e. those in which contradictions can be proved (or in which the logic is *dialetheist*) without leading to inconsistency in the sense that all statements follow; the underlying logic for these must thus exclude at a minimum, *ex falso quodlibet*, but there are also other complicating restrictions that need to be made (cf., e.g., Graham Priest (2002)). So far as I know, it has not been determined whether such logics account for "sustained ordinary reasoning", not only in everyday discourse but also in mathematics and the sciences. If they do, they deserve serious consideration as a possible route under (2°). In any case, I have chosen to base my systems on ordinary classical two-valued logic, which certainly meets this criterion. That leaves the route (3°); and here the obvious principle needing restriction is what Tarski called the material adequacy condition for truth, namely the T-scheme, according to which T(#A) is equivalent to A for all sentences A, where #A names A in some way or other. (For simplicity, in the following I will omit the '#' sign within the T predicate.) The restriction of the T-scheme taken here is motivated by the following points:

(1) I agree with Bertrand Russell (1908) that every predicate has a *domain of significance*, and it makes sense to apply the predicate only to objects in that domain. In the case of truth, that domain D consists of the sentences that are *meaningful* and *determinate*, i.e. have a definite truth value, true or false. D includes various but not necessarily all grammatically correct sentences that involve the notion of truth itself.[2]

[2] Gödel seemed to be in accord with this idea in his article, "Russell's mathematical logic" (1944), p.149: "It should be noted that the theory of types brings in a new idea for the solution of the paradoxes, especially suited to their intensional form. It consists in blaming the paradoxes not on the axiom that every propositional function defines a concept or class, but on the assumption that every concept gives a meaningful proposition, if asserted for any arbitrary object or objects as arguments." This is the point of departure for Reinhardt (1986), 221 ff; Reinhardt uses the word 'significant', with predicate symbol S(x),

(2) Some authors (e.g., Kripke 1975) consider sentences like the Liar to be meaningful, though they do not have a determinate truth value (nor, for that matter, does the Truth Teller). My own view is that the Liar is not meaningful, but in order to avoid confusion and to allow for such differences of opinion, have added the modifier 'determinate'. In any case, T(A) implies D(A) for each sentence A.

(3) Thus the restriction of the T-scheme should take the form,

$$D(A) \rightarrow (T(A) \leftrightarrow A).$$

(4) Taking the logic of truth to be classical, and writing F(A) for T(¬A) it follows that

$$D(A) \rightarrow (T(A) \vee F(A))$$

for each sentence A. But since both $T(A) \rightarrow D(A)$ and $F(A) \rightarrow D(A)$ by (1), we must have

$$D(A) \leftrightarrow (T(A) \vee F(A))$$

for each A.

(5) Though this serves to identify D in terms of T, the conditions on D should be prior to those on T, i.e. determinate meaningfulness is prior to truth.[3] First of all, D is closed under the propositional operations and quantifiers of the first-order predicate calculus (where a formula is taken to belong to D if all its substitution instances by meaningful terms belongs to D.) In the opposite direction, a sentence is meaningful only if all of its parts are meaningful. But that does not hold for determinateness by itself; for example, a disjunction $A \vee B$ can be considered to have the truth value true if A is true even when B has no truth value. On the other hand, if $A \vee B$ is both meaningful and determinate, each of A and B must be meaningful. Though that does still not force both A and B to be determinate when at least one of them is true, it is reasonable to require that if we are to have a general rule for determining the truth value of $A \vee B$ that works internally to D. And for those who, like me, think that meaningfulness and determinateness coincide, this condition is automatic. At any rate, the closure conditions for D are assumed here to be invertible, and are thus of a biconditional, or *strongly compositional* form when combined with the closure conditions.

---

where I use 'determinate meaningfulness' with predicate symbol D(x). The terminology 'determinate meaningfulness' with predicate symbol M(x) is used by McDonald (2000).

[3] In broad terms, that was also the program of Reinhardt (1985), leading to a reformulation **KS** of **KF** in terms of axioms for significance followed by those for truth.

(6) If L_0 is a language of which we recognize that each sentence A satisfies D, and $\mathbf{S}_0$ is a theory in L_0 that we recognize informally to be correct, then we should be able to prove in an extension **S** of $\mathbf{S}_0$ for D and T that each theorem A of $\mathbf{S}_0$ satisfies T. *Moreover*, the statement to that effect should itself satisfy T.

An essential difference of what is done here from (F 84) is that in the latter, the conditions on D are *posterior* to those on T, not *prior* to them as required by point (5).

In the next section I shall produce a specific system **DT** whose formulation is motivated by (1)-(6). The consistency of **DT** is proved in sec. 3 by construction of a model M. The paper concludes with a discussion in sec. 4, including assessment of how far **DT** goes to meeting the criteria of Leitgeb (2007).

2. **The system DT** To see how close we can come to following out the preceding ideas in a specific setting, let L_0 be the language of arithmetic (**PA**, our $\mathbf{S}_0$), and let $L = L_0(T)$ be the extension of L_0 by a new unary predicate T. In this case, #A is the numeral of the Gödel number of A and we write T(A) for T(#A) for each sentence A of L. Let F(x) be $T(\neg.x)$ and D(x) be $T(x) \vee F(x)$. The system **DT** in the language L is designed to satisfy the following conditions:

(i) its logic is that of the classical first order predicate calculus;

(ii) **DT** has a model in an expansion of the standard model for **PA**; we thus require that **PA** $\subseteq$ **DT** and **DT** includes full induction on the natural numbers for all formulas of L;

(iii) D provably satisfies the strongly compositional (i.e., biconditional) conditions meeting the requirement (5) above;

(iv) for x satisfying D, T(x) provably satisfies the usual recursive defining conditions;

(v) **DT** proves $D(A) \rightarrow (T(A) \leftrightarrow A)$ for each A of L;

(vi) in accordance with (6), **DT** proves $T(\forall x(Sent_0(x) \wedge Prov_{PA}(x) \rightarrow T(x))$, where $Sent_0(x)$ expresses that x is (the Gödel number of) a sentence of L_0.

As examples for (iii), (iv) I mean that **DT** proves general statements of the following kinds (that will be elaborated below):[4]

[4] The dot notation with logical symbols in the following serves to indicate the corresponding primitive recursive operations on Gödel numbers; 'num' is the operation that takes a natural number and returns the corresponding numeral in L_0. $Sent_L(x)$ expresses that x is the Gödel number of a sentence of L. Other formalized predicates of metamathematical notions are explained similarly.

(a) $D(T(num(x))) \leftrightarrow D(x)$, and $D(x) \rightarrow [T(T(num(x)) \leftrightarrow T(x)]$;

(b) $D(\neg.x) \leftrightarrow D(x)$, and $D(x) \rightarrow [T(\neg.x) \leftrightarrow \neg T(x)]$,

(c) $D(x \vee. y) \leftrightarrow D(x) \wedge D(y)$, and $D(x \vee. y) \rightarrow [T(x \vee. y) \leftrightarrow T(x) \vee T(y)]$,

(d) $D(\forall z.\ x) \leftrightarrow \forall y\ D(sub(num(y), z, x))$ if Var(z), and

$D(\forall z.\ x) \rightarrow [T(\forall z.\ x) \leftrightarrow \forall y\ T(sub(num(y), z, x)]$.

We shall also want that every sentence of L_0 satisfies D(x). It is then shown by induction in **DT** that all numerical substitution instances of provable formulas of **PA** are true, and hence that **DT** proves $\forall x(Sent_0(x) \wedge Prov_{PA}(x) \rightarrow T(x))$. But it does not follow that that sentence is provably true. In fact, if $\rightarrow$ is defined as usual in terms of $\neg$ and $\vee$, there is a problem about that. For, take λ to be a "liar" sentence, i.e. one for which **DT** proves $\lambda \leftrightarrow \neg T(\lambda)$; then we have $\neg D(\lambda)$, since otherwise we would have $(T(\lambda) \leftrightarrow \lambda)$ and arrive at a contradiction.. It then follows from (a) that we have $\neg D(T(\lambda))$. However, this will lead to a problem for (vi) with (b)-(d) if we identify $A \rightarrow B$ with $(\neg A \vee B)$. Arguing informally, to see that the sentence in (vi) satisfies the T predicate, we will want to show it satisfies the D predicate and then apply (d). Thus if we make that identification we will need to show that for each numeral n,

$D(\neg(Sent_0(n) \wedge Prov_{PA}(n)) \vee T(n))$ holds.

But this will require that D(T(n)) must hold for each n and so, in particular, we must have $D(T(\lambda))$, which is excluded.

To get around this problem, we *do not identify* $A \rightarrow B$ with $(\neg A \vee B)$ but take it as a separate basic propositional operation, so as to treat the D predicate applied to conditionals in a different way, namely:

(e) $D(x \rightarrow. y) \leftrightarrow D(x) \wedge (T(x) \rightarrow D(y))$, and $D(x \rightarrow. y) \rightarrow [T(x \rightarrow.y) \leftrightarrow (T(x) \rightarrow T(y))]$.

However, the *logic* of $\rightarrow$ is unchanged, as is the determination of T applied to conditionals; it is only the determination of the D predicate applied to conditionals that is modified.[5] Note that we do not in this case meet the requirement (5) above in full.[6]

[5] Condition (e) is analogous to Aczel's (1980) treatment of $\rightarrow$ in Frege structures; the relation of **DT** to the defining conditions for Frege structures is explained below.

[6] In the discussion following the presentation of this material at the Banff workshop it was argued by both Yiannis Moschovakis and Stewart Shapiro that the condition (e) on D is natural since we don't care whether $D(A \rightarrow B)$ holds when it is determined that T(A) doesn't hold.

Let $\mathbf{PA}_L$ be the extension of **PA** by all instances of the induction scheme in L. The system **DT** is now specified to consist of the following three groups of axioms:

I. $\mathbf{PA}_L$

II. (i) $\forall x[\text{At-Sent}_0(x) \rightarrow D(x)]$

(ii) $\forall x[\, D(T.(num(x)) \leftrightarrow D(x)]$

(iii) $\forall x[\text{Sent}_L(x) \rightarrow (D(\neg.x) \leftrightarrow D(x)]$

(iv) $\forall x \forall y[\, \text{Sent}_L(x) \wedge \text{Sent}_L(y) \rightarrow (D(x \vee .y) \leftrightarrow D(x) \wedge D(y)]$,

(v) $\forall x \forall y[\text{Sent}_L(x) \wedge \text{Sent}_L(y) \rightarrow (D(x \rightarrow .y) \leftrightarrow D(x) \wedge (T(x) \rightarrow D(y))\ ]$,

(vi) $\forall x, z\ [\text{Var}(z) \wedge \text{Sent}_L(\forall z.\ x)) \rightarrow (D(\forall z.\ x) \leftrightarrow \forall y\ D(\text{sub}(num(y), z, x)]$.

III. (i) for each atomic formula $R(x_1,\ldots,x_k)$ of L_0,

$\forall x_1 \ldots \forall x_k[\, T(R.(num(x_1),\ldots,num(x_k)) \leftrightarrow R(x_1,\ldots,x_k)\ ]$

(ii) $\forall x[\, D(x) \rightarrow (\, T(T.(num(x)) \leftrightarrow T(x)\ )\ ]$

(iii) $\forall x\ [\text{Sent}_L(x) \wedge D(x) \rightarrow (T(\neg.x) \leftrightarrow \neg T(x))\ ]$

(iv) $\forall x \forall y[\, \text{Sent}_L(x) \wedge \text{Sent}_L(y) \wedge D(x \vee .y) \rightarrow (T(x \vee .y) \leftrightarrow T(x) \vee T(y))\ ]$.

(v) $\forall x \forall y[\, \text{Sent}_L(x) \wedge \text{Sent}_L(y) \wedge D(x \rightarrow .y) \rightarrow (T(x \rightarrow .y) \leftrightarrow (T(x) \rightarrow T(y))\ ]$.

(vi) $\forall x, z\ [\, \text{Var}(z) \wedge \text{Sent}_L(\forall z.\ x)) \wedge D(\forall z.\ x) \rightarrow$

$(T(\forall z.\ x) \leftrightarrow \forall y\ T(\text{sub}(num(y), z, x))\ ]$.

Remark. For those who know the work of Aczel (1980), it is seen that the axioms for D, T are similar to those for proposition and truth in Frege structures, op. cit. p. 37. One essential difference is that Aczel only imposes closure conditions on the notion of proposition; that would correspond to weakening the axioms in group II for D by replacing '$\leftrightarrow$' throughout by '$\leftarrow$'. A second essential difference is that Aczel does not have conditions for propositions of the form T(s), and thus none corresponding to our axioms II(ii) and III(ii). An inessential difference is that Aczel's notion of Frege structure is not given as an axiomatic theory.[7] Another difference lies in the basic framework; here it is arithmetic, while for Frege structures it is the λ-calculus. The latter allows for more general interpretations; further work on systems like **DT** might usefully

[7] Beeson (1985), pp. 410ff, provides an axiomatic version **F** of Frege structures.

incorporate similar features. Finally, the proof of consistency of **DT** given in the next section is somewhat different from Aczel's proof of the existence of Frege structures.

Theorem 1. **DT** proves the following:

(i) D(A) for each sentence A of L_0.

(ii) $D(A) \rightarrow [T(A) \leftrightarrow A]$ for each sentence A of L.

(iii) $T(\forall x[Sent_0(x) \wedge Prov_{PA}(x) \rightarrow T(x)])$.

Proof. For (i), one first proves more generally for each formula $A(x_1,\ldots,x_k)$ of L_0 that $D(A(num(x_1),\ldots,num(x_k))$ is provable in **DT**, by induction on the formation of A. (ii) is handled similarly for formulas A of L. To prove (iii), let us reason informally in **DT**. We first show by induction on proofs that for any formula A of L_0 with free variables $x_1,\ldots,x_k$ that is provable in **PA**, we have $T(A(num(x_1),\ldots,num(x_k))$.[8] Hence if A is any sentence of L_0 provable from **PA** then T(A), i.e. we have $\forall x[Sent_0(x) \wedge Prov_{PA}(x) \rightarrow T(x)]$. Now to show that that sentence—call it B—is true in the sense that T(B) holds, it is sufficient by(ii) to show that D(B) holds. Informally, this comes down to showing that for each n, $D(Sent_0(n) \wedge Prov_{PA}(n) \rightarrow T(n))$ holds, and that is equivalent to showing that $D(C) \wedge [T(C) \rightarrow D(T(n))]$, where C is $(Sent_0(n) \wedge Prov_{PA}(n))$. C is a sentence of L_0 so D(C) holds and, moreover, $T(C) \leftrightarrow C$. Now we already know that $C \rightarrow T(n)$, and of course $T(n) \rightarrow D(T(n))$, so that completes the argument.

3. Construction of a standard model for DT. Let $M_0 = (N,\ldots)$ be a standard model for **PA** in the language L_0. We shall construct an expansion of M_0 to a model M for **DT** in the language $L = L_0(T)$. M will in turn be obtained from a certain three-valued model M*, where the values lie in the set $\mathbf{3} = \{t, f, u\}$. In evaluating compound sentences of L in M*, built up by $\neg$, $\vee$, $\rightarrow$, and $\forall$, we make use of weak Kleene semantics (slightly modified in the case of $\rightarrow$) for evaluating corresponding operations on **3**. Relying on context to avoid ambiguity, we also use the symbols $\neg$, $\vee$, $\rightarrow$, for the corresponding operations on **3**, while we use $\prod$ for the infinitary operation corresponding to $\forall$. For $a \in \mathbf{3}$ write D(a) for (a = t or a = f).

[8] Note that the verification of this for all instances of the induction scheme in PA may be provided by a single instance of the induction scheme in PA_L.

Definition 2.

(i) $D(\neg a)$ iff $D(a)$; if $D(\neg a)$ then $(\neg a) = t$ iff $a = f$, else $(\neg a) = u$.

(ii) $D(a \vee b)$ iff $D(a)$ and $D(b)$; if $D(a \vee b)$ then $(a \vee b) = t$ iff $(a = t$ or $b = t)$;
else $(a \vee b) = u$.

(iii) $D(a \rightarrow b)$ iff $D(a)$ & $(a = f$ or $D(b))$; if $D(a \rightarrow b)$ then
$(a \rightarrow b) = t$ iff $(a = f$ or $b = t)$; else $(a \rightarrow b) = u$.

(iv) $D(\prod\{a_i : i \in I\})$ iff $D(a_i)$ for each $i \in I$; if $D(\prod\{a_i : i \in I\})$ then
$\prod\{a_i : i \in I\} = t$ iff for each $i \in I$, $a_i = t$; else $\prod\{a_i : i \in I\} = u$.

Define $(a \wedge b) = \neg(\neg a \vee \neg b)$. Then we have: $D(a \wedge b)$ iff $D(a)$ and $D(b)$; and if $D(a \wedge b)$ then $(a \wedge b) = t$ iff $a = b = t$, else its value is u. Similarly, define $\sum\{a_i : i \in I\} = \neg\prod\{\neg a_i : i \in I\})$. Then we have: $D(\sum\{a_i : i \in I\})$ iff $D(a_i)$ for each $i \in I$; and if $D(\sum\{a_i : i \in I\})$ then $\sum\{a_i : i \in I\} = t$ iff $a_i = t$ for some $i \in I$, else its value is u. We cannot eliminate $\rightarrow$ in favor of $\neg$ and $\vee$.

Lemma 3. Each of $\neg$, $\vee$, $\rightarrow$ and $\prod$ is monotonic on the reflexive ordering of {t, f, u} with $u \leq t$, $u \leq f$.

Proof. Immediate for all the operations except $\rightarrow$. To prove it for that, consider any a, b, a', b' in **3** with $a \leq a'$ and $b \leq b'$; to show $(a \rightarrow b) \leq (a' \rightarrow b')$. If both D(a) and D(b) this is trivial. Suppose a = u; then not $D(a \rightarrow b)$ so $(a \rightarrow b) = u$, and that is $\leq$ any value. Thus we may assume a = t or a = f, and b = u. Since $(t \rightarrow u) = u$, that is $\leq$ any value. The final case is $(f \rightarrow u)$, which = t. Then whatever b' is, we have $(f \rightarrow b') = t$, too.

In consequence of this lemma, the Kripke (1975) style construction yields an expansion of M_0 to a **3**-valued structure $M^* = (M_0, T^*)$, where T^* is the least fixed point **3**-valued predicate under the evaluation of each sentence A of L according to the rules given by Definition 2, together with the requirement that T(A) evaluates the same as A. Note that Kripke used strong Kleene three-valued semantics in his construction, though he pointed out that it works just as well for monotonic operations more generally. Thus we have the

following theorem, where we write v(A) for the value of A in **3** given by the semantics of M*.[9]

Theorem 4. There is a **3**-valued model M* of L given by an assignment v(A) in {t, f, u} to each sentence A of L satisfying the following conditions.

(i)(a) if A is an atomic sentence of L_0 then v(A) = t or v(A) = f and v(A) = t iff $M_0 \models A$.

(i)(b) if s is a closed term and s denotes a sentence A in L, then v(T(s)) = v(A), otherwise v(T(s)) = f.

(ii) $v(\neg A) = \neg v(A)$

(iii) $v(A \vee B) = v(A) \vee v(B)$

(iv) $v(A \rightarrow B) = v(A) \rightarrow v(B)$.

(v) $v(\forall x\, A(x)) = \prod\{v(A(n)) : n \in N\}$.

Following the approach of Aczel utilized in (F 84) §11, M* is then converted into a 2-valued model $M = (M_0, T)$ of L so that the predicate T holds of n in M iff v(T(n)) = t in M*. The following then specifies satisfaction in M in the standard way.

Definition 5.

(i) If A is an atomic sentence of L then $M \models A$ iff v(A) = t in M*

(ii) $M \models \neg A$ iff not $M \models A$

(iii) $M \models A \vee B$ iff $M \models A$ or $M \models B$

(iv) $M \models A \rightarrow B$ iff not $M \models A$ or $M \models B$

(v) $M \models \forall x\, A(x)$ iff $M \models A(n)$ for each $n \in N$.

Lemma 6.

(i) $M \models T(A)$ iff v(A) = t

(ii) $M \models F(A)$ iff v(A) = f

(iii) $M \models D(A)$ iff D(v(A))

(iv) $M \models D(T(A))$ iff D(v(A)).

[9] In (F 84) I wrote $\| A \|$ for v(A).

Corollary 7.

(i) For any L_0 sentence A: $M \models D(A)$

(ii) For any L sentence A: $M \models D(T(A)) \leftrightarrow D(A)$

(iii) For each L sentence A: $M \models D(\neg A) \leftrightarrow D(A)$

(iv) For any L sentences A, B: $M \models D(A \vee B) \leftrightarrow D(A) \wedge D(B)$

(v) For any L sentences A, B:

$M \models D(A \rightarrow B)$ iff $M \models D(A) \wedge (T(A) \rightarrow D(B))$.

(vi) For any L sentence $\forall x\, A(x)$:

$M \models D(\forall x\, A(x))$ iff for each n, $M \models D(A(n))$.

Theorem 8.

(i) If A is a sentence of L_0 then $M \models A$ iff $M_0 \models A$.

(ii) For each sentence A of L, if $M \models D(A)$ then $M \models A$ iff $v(A) = t$.

(iii) For each sentence A of L, $M \models D(A) \rightarrow (T(A) \leftrightarrow A)$.

(iv) For each sentence A of L, $M \models T(A) \rightarrow A$.

Proofs (i) is immediate by definition of M.

(ii) is proved by induction on the logical complexity of A. It holds for atomic A by Defn. 5(i). Suppose it holds for A and that $D(\neg A)$ holds. Then D(A) holds, so by induction we have $M \models A$ iff $v(A) = t$; thus t.f.a.e.: $M \models \neg A$, $v(A) \neq t$, $v(A) = f$, and $v(\neg A) = t$. Suppose it holds for A, B and suppose $D(A \vee B)$ holds in M. Then by Lemma 6, both D(A) and D(B) hold in M. By induction, $M \models A$ iff $v(A) = t$ and $M \models B$ iff $v(B) = t$, so $M \models (A \vee B)$ iff $v(A \vee B) = t$. Suppose it holds for A, B and suppose $D(A \rightarrow B)$ holds in M. Then D(A) holds and $(T(A) \rightarrow D(B))$ holds. To show that $A \rightarrow B$ holds in M iff $v(A \rightarrow B) = t$, i.e. iff $D(v(A))$ and ($v(A) = f$ or $v(B) = t$). Since D(A) holds, we have $D(v(A))$, and $M \models A$ iff $v(A) = t$ by induction hypothesis. Thus $M \models (A \rightarrow B)$ iff ($v(A) = t$ implies $M \models B$), and we are reduced to showing that $M \models B$ iff $v(B) = t$ when $v(A) = t$. But in that case $v(T(A)) = t$ as well, and $M \models T(A)$ by definition of M. Since $T(A) \rightarrow D(B)$ holds in M, it follows that $M \models D(B)$; thus we can now apply the induction hypothesis to B, which is exactly what is required to complete this step. Finally, suppose

the induction hypothesis holds for A(n) for each n, and suppose D(∀x A(x)) holds in M, then t.f.a.e. : M |= ∀x A(x); for all n ∈ N, M |= A(n); for all n ∈ N, v(A(n)) = t; v(∀x A(x)) = t.

(iii) is a corollary of (ii) since if D(A) holds in M, M |= A iff v(A) = t, which is equivalent to v(T(A)) = t, and hence to M |= T(A).

(iv) is then immediate, since T(A) implies D(A) by definition of D(A) as T(A) ∨ F(A).

Theorem 9. M is a model of **DT**.

Proof. Since M = (M_0, T) is standard for the natural numbers, the axioms for **DT** are all true in M by the preceding results.

Conjectures

(C1) The proof-theoretic strength of **DT** is the same as that of **RA**(< ε_0), Ramified Analysis in levels up to the ordinal ε_0. I expect a proof of this would follow the methods of (F 91) pp. 25-30. For the upper bound, one makes use of the fact observed by Aczel that the existence of a fixed point for a positive arithemetical inductive definition can be established in the system $\sum^1_1$-**AC**, whose strength was shown by Harvey Friedman to be the same as that of **RA**(< ε_0). The fixed point statement is used to produce a three-valued model M* satisfying Theorem 4 above; the construction of M from M* is arithmetical. For the lower bound, one uses the fact that transfinite induction up to α for each $\alpha < \varepsilon_0$ can be established in **DT** for all formulas of L. See loc. cit. for more details.[10]

(C2) Consider a parametric form **DT**(P) of **DT** like that of Ref*(**PA**(P)) in (F 91), obtained by adding a predicate parameter P, relativizing T to P and adding a rule of substitution. I conjecture that the proof-theoretic strength of such a system **DT**(P) is the same as that of **RA**(< Γ_0), Ramified Analysis in levels up to the least impredicative ordinal Γ_0. A proof would be similar to that for determining the strength of Ref*(**PA**(P)) in (F 91), pp. 30 ff.

[10] Since writing the above I have been informed by Kentaro Fujimoto that he has found an interpretation of **DT** in **KF**, thus confirming the conjectured upper bound. Also, Thomas Strahm assures me that the argument sketched in the text does indeed go through for both the upper and lower bound.

Question: is there a natural non-parametric extension of **DT** with the same strength as predicative analysis?

4. Discussion. Before getting into broader issues, let's look at how the usual statements that lead to contradictions with the unrestricted T-scheme are accounted for in **DT**.

1. The liar. Let λ be a sentence of L such that **DT** $\vdash \lambda \leftrightarrow F(\lambda)$. Then **DT** proves $\neg D(\lambda)$, for otherwise we would have λ equivalent to $T(\lambda)$ and thence to $\neg F(\lambda)$. Thus we cannot obtain the usual contradiction from the T-scheme in its D-restricted form (Theorem 1 (ii)). Similarly if we take λ to be such that **DT** $\vdash \lambda \leftrightarrow \neg T(\lambda)$.

2. The strengthened liar. Let σ be a sentence of L such that **DT** $\vdash \sigma \leftrightarrow \neg D(\sigma) \vee F(\sigma)$. Reasoning informally, if $D(\sigma)$ then σ is equivalent to $F(\sigma)$, so by 1, we have a contradiction. Hence $\neg D(\sigma)$. Informally then, σ is true though not determinate in the sense of satisfying D. But on our theory, truth as given by the predicate T does not hold of σ or the r.h.s. of its defining equivalence.

3. Alternative strengthened liar. Let σ^* be a sentence of L such that **DT** $\vdash \sigma^* \leftrightarrow (D(\sigma^*) \rightarrow F(\sigma^*))$. Then again we conclude $\neg D(\sigma^*)$. Now the r.h.s can't satisfy T, since that requires $D(D(\sigma^*))$, which is equivalent to $D(\sigma^*)$.

Various criteria have been proposed for consistent formal theories of truth that contain their own truth predicate. To my mind, the best articulation of these is a recent one due to Hannes Leitgeb (2007).[11] He sets down eight of these, each of which has a plausibility in its own right, and various of which are ordinarily taken for granted, but the combination of all of which cannot simultaneously be realized on pain of inconsistency. That is why Leitgeb calls his piece, "What theories of truth should be like (but cannot be)". His eight criteria (a-h) are as follows.

a. Truth should be expressed by a predicate (and a theory of syntax should be available).

[11] But see also Sheard (2002). Another interesting discussion of criteria is to be found in McDonald (2000).

b. If a theory of truth is added to mathematical or empirical theories, it should be possible to prove the latter true.

c. The truth predicate should not be subject to any type restrictions.

d. T-biconditionals [in the T-scheme] should be derivable unrestrictedly.

e. Truth should be compositional.

f. The theory should allow for standard interpretations.

g. The outer logic and the inner logic should coincide.

h. The outer logic should be classical.

I won't repeat Leitgeb's elaborations of what these mean except for g, which he explains as follows:

> When truth theorists refer to the "outer" and the "inner" logic of a theory of truth, what they mean is that the logical laws in such theories can show up in two different contexts: outside of applications of '*Tr*' and inside of such contexts. E.g., there are consistent theories of truth in which both sentences of the form 'A or not A' and 'not *Tr*('A or not A')' are derivable. While the former are instances of the classical law of excluded middle, the latter deny instances of excluded middle in the context of the truth predicate. Accordingly, although the outer logic of the theory might be genuinely classical, its inner logic certainly is not. This is in contrast with e.g. Tarski's theory, which is an example of a theory of truth for which the outer and the inner logic coincide (in either case, classical logic).

Thus, for example, the outer logic of **KF** is classical, while its inner logic is that of strong Kleene 3-valued logic, as in Kripke's model.[12] The situation is a little different for the system **DT**, whose outer logic is classical while its inner logic is classical only for sentences satifying the D predicate, since e.g. it proves $\neg T(\lambda \vee \neg\lambda)$. Note that d, the unrestricted T-scheme, implies g. In Leitgeb's view, the importance of condition g is that

[12] Halbach and Horsten (2006) have devised a modification of the system **KF** to make the outer and inner logic coincide with strong Kleene 3-valued logic.

> whatever reasons there might be for preferring one logic over another, if they apply to linguistic contexts outside the applications of truth predicates, why should they not equally apply to contexts within such applications? Every discrepancy between the outer and the inner logic of a theory of truth would indicate that our calling a sentence true somehow changes the logic that governs our understanding of this sentence. This is definitely questionable. Hence such discrepancies are—ceteris paribus—to be voted out.[13]

I would add a further criterion to a-h, namely that the logic of the ambient metatheory used to establish consistency of one's theory should be the same as the logic basic to that theory (i.e., its outer logic); this holds for both the systems **KF** and **DT**. It does not hold of systems obtained by restricting its basic logic somehow, as is the case, for example, with the paraconsistent logics of Priest (2002) or the logic of Field (2003).[14]

Going down Leitgeb's list, **DT** meets the criteria directly or with restriction as follows.

a. Met.

b. Met for the specific case of **PA** (as a working example); it should be of interest to extend this so as to be applicable to suitable empirical theories as well.

c. Met.

d. Met only for those sentences satisfying the D predicate.

e. Met under the same restriction.

f. Met.

g. Met only to the extent that the inner logic is classical for sentences satisfying the D predicate.

h. Met.

---

[13] The distinction between outer logic and inner logic has been more or less explicit in critical discussions of formal theories of truth over the years. Reinhardt's work (1986) was a sustained but ultimately unsuccessful effort to resolve the discrepancy.

[14] Field (2003) augments strong Kleene logic by a new conditional $\Rightarrow$ for which the law of importation fails; the T-scheme is formulated with the associated biconditional $\Leftrightarrow$. See the discussion of Field's system near the end of Leitgeb (2007).

Remarks

1. The failure of g in general for **DT** is a *prima-facie* mark against it, if we take **DT** to be a first-class citizen in its own right. A fall-back position would be to treat it instrumentally, in the spirit of (F 84).[15] But for that one would want to have more than consistency, or even conservation over **RA**($< \varepsilon_0$) (assuming the conjecture (C1) at the end of sec. 3 above is right). It is not clear to me just what that "more" should be.

2. Continuing the instrumentalist tack, one can interpret a type-free theory of sets in **DT** by taking $y \in x$ to mean that x is the Gödel number of a formula A with one free variable y such that T(A(num(y))). This is type-free in the sense that various instances of self-membered sets can be produced; how far this could be made useful, for example in the sense of applications to category theory, remains to be seen; see my papers (2004), (2006) for desiderata.

3. The restrictions in d and e seem to me to be more or less compelling, for the reasons given in sec. 1. As to g, the provability in **DT** of sentences $\neg$T(A) for which A is provable might be regarded as "unintended consequences" or "anomalies" or "little monsters". In a way this is analogous to other situations in mathematics. For example, to develop a good theory of integration, Lebesgue introduced his theory of measure; that has many excellent properties but also the unintended consequence that there are non-measurable sets whose existence does not affect the positive applications of the theory. Another example is the existence of space-filling curves as a consequence of a good theory of continuous mappings formulated in purely topological terms.

4. If the eventual aim is a theory of truth without such unintended consequences, perhaps a two stage affair can work: a *theory of (determinate) meaningfulness* followed by a *theory of truth*. In the first stage one would ascertain (presumably in a non-effective way) which sentences of L are meaningful and have determinate truth value. In the second stage one would introduce a new kind of variable ranging over just those sentences and apply the predicate T *only* to terms of that kind. I have made efforts in this

[15] Reinhardt (1986) also urged an instrumentalist view of such theories, in analogy to Hilbert's program.

direction that have so far been unsuccessful, but I am optimistic that something like this can be done, and moreover in a clean and informally convincing way.[16]

References

P. Aczel (1980), Frege structures and the notions of proposition, truth and set, in *The Kleene Symposium* (J. Barwise, et al., eds.), North-Holland, Amsterdam, 31-59.

M. J. Beeson (1985), *Foundations of Constructive Mathematics*, Springer-Verlag, Berlin.

A. Cantini (1989), Notes on formal theories of truth, *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 35, 97–130.

S. Feferman (1984), Toward useful type-free theories I, *J. Symbolic Logic* 49, 75-111.

_____ (1991), Reflecting on incompleteness, *J. Symbolic Logic* 56, 1-49.

_____ (1996), Gödel's program for new axioms: Why, where, how and what?, in *Gödel '96* (P. Hájek, ed.), *Lecture Notes in Logic* 6, 3-22.

_____ (2004), Typical ambiguity: trying to have your cake and eat it too, in *One Hundred Years of Russell's Paradox* (G. Link, ed.), Walter de Gruyter, Berlin, 131-151.

_____ (2006), Enriched stratified systems for the foundations of category theory, in *What is Category Theory?* (G. Sica, ed.), Polimetrica, Milano, 185-203.

S. Feferman and T. Strahm (2000), The unfolding of non-finitist arithmetic, *Annals of Pure and Applied Logic* 104, 75-96.

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

K. Gödel (1944), Russell's mathematical logic, in *The Philosophy of Bertrand Russell* (P. A. Schilpp, ed.), Library of Living Philosophers, Northwestern, Evanston, 123-153; reprinted in Gödel (1990), 119-141.

K. Gödel (1990), *Collected Works, Vol. II. Publications 1938-1974* (S. Feferman, et al., eds.), Oxford University Press, NY.

V. Halbach (2007), Axiomatic theories of truth, *Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), URL = < http://plato.stanford.edu/entries/truth-axiomatic>.

V. Halbach and L. Horsten (2006), Axiomatizing Kripke's theory of truth, *J. Symbolic Logic* 71, 677–712.

S. Kripke (1975), Outline of a theory of truth, *J. of Philosophy* 72, 690-716.

H. Leitgeb (2007), What theories of truth should be like (but cannot be), *Philosophy Compass* 2 (2), 276-290.

B. E. McDonald (2000), On meaningfulness and truth, *J. Philosophical Logic* 29, 433-482.

V. McGee (1991), *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*, Hackett, Indianapolis.

G. Priest (2002), Paraconsistent logic", in *Handbook of Philosophical Logic* (2nd Edn.), Vol. 6 (D. Gabbay and F. Guenthner, eds.), Kluwer, Dordrecht, 287–393.

[16] The work of McDonald (2000) may be considered to be a step in this direction. He provides a semantics for (determinate) meaningfulness and truth in transfinite stages M_α and T_α, where the definition of the former at each stage α precedes the definition of the latter; however, M and T are not separated in general.

W. N. Reinhardt (1985), Remarks on significance and meaningful applicability, in *Mathematical Logic and Formal Systems* (L. Paulo de Alcantara, ed.), *Lecture Notes in Pure and Applied Mathematics* 94, 227-242.

W. N. Reinhardt (1986), Some remarks on extending and interpreting theories,with a partial predicate for truth, *J. Philosophical Logic* 15, 219–51.

B. Russell (1908), Mathematical logic as based on the theory of types, *Amer. J. Mathematics* 30, 222-262; reprinted in *From Frege to Gödel. A source book in mathematical logic, 1879-1931* (J. van Heijenoort, ed.), 1967, Harvard Univ. Press, Cambridge, 150-182.

M. Sheard (1994), A guide to truth predicates in the modern era, *J. Symbolic Logic* 59, 1032–54.

M. Sheard (2002), Truth, provability, and naive criteria, in *Principles of truth* (V. Halbach and L. Horsten, eds.), Dr. Hänsel-Hohenhausen, Frankfurt a.M., 169-181.