[Paper for the Ohio State University conference in honor of the 60th birthday of Harvey Friedman]

FRIEDMAN AND THE AXIOMATIZATION OF KRIPKE'S THEORY OF TRUTH

ABSTRACT

What is the simplest and most natural axiomatic replacement for the set-theoretic definition of the minimal fixed point on the Kleene scheme in Kripke's theory of truth? What is the simplest and most natural set of axioms and rules for truth whose adoption by a subject who had never heard the word "true" before would give that subject an understanding of truth for which the minimal fixed point on the Kleene scheme would be a good model? Several axiomatic systems, old and new, are examined and evaluated as candidate answers to these questions, with results of Harvey Friedman playing a significant role in the examination.

John P. Burgess
Department of Philosophy
Princeton University
Princeton, NJ 08544-1006 USA
jburgess@princeton.edu

*The author is grateful to Andrea Cantini, Graham Leigh, Leon Horsten, and Michael Rathjen for useful comments on earlier drafts of this paper, and to Jeremy Avigad for background information on proof-theoretic matters.

1

Small though it is, the area of logic concerned with axiomatic theories of truth is large enough to have two distinguishable sides. These go back to contrasting early reactions of two eminent logicians to Saul Kripke's "Outline of a Theory of Truth" [1975]. One side originates with Harvey Friedman, who first wrote Kripke in the year of the publication of the "Outline", but whose published contributions are contained in a joint paper with Michael Sheard from over a decade later, Friedman and Sheard [1987]. (There was also a sequel, Friedman and Sheard [1988], but I will not be discussing it.) The questions raised in that paper are these: First, which combinations of naive assumptions about the truth predicate are consistent? Second, what are the proof-theoretic strengths of the consistent combinations?

In the Friedman-Sheard paper, combinations of items from a menu of a dozen principles are added to a fixed base theory that includes first-order Peano arithmetic PA. A variety of model constructions are presented to show various combinations consistent, and a number of deductions to show various other combinations inconsistent, and complete charts of the status of all combinations worked out. There turn out to be nine maximal consistent sets.

In a portion attributed in the paper to Friedman alone (§7), two sample results on proof-theoretic strength are presented, showing one combination very weak and another very strong. Later additional results on proof-theoretic strength were obtained by a number of workers, and most recently Graham Leigh and Michael Rathjen [forthcoming] have finished the job, so that we now have a complete determination of the proof-theoretic strengths of all nine maximal consistent sets.

Though the questions addressed in Friedman's work are purely mathematical, and the paper with Sheard explicitly declares its philosophical neutrality, the notion of truth is so philosophically fraught that one naturally expects some of the formal results will turn out to have some bearing on questions of interest to philosophers. This expectation is not

disappointed, and I will be making use of Friedman's proofs of both his sample results in the course of this paper.

2

I follow the example of Friedman and Sheard by describing in advance the base language and theory to be considered, and in listing and naming the various candidate principles of truth. (See the table of PRINCIPLES OF TRUTH.) The base language will be that of arithmetic with a truth predicate T. Formulas not involving the new predicate are called *arithmetical*. Sometimes it will be convenient to have also a falsehood predicate F, where *falsehood* is truth of the negation (as *denial* is assertion of the negation, and *refutation* is proof of the negation), while the negation of truth is *untruth*. F need not be thought of as a primitive but may be thought of as defined. (Some truth principles that are nontrivial when it is taken as primitive become trivial when it is taken as defined.) T(x) literally means "x is the code number for a true sentence". The coding of sentences and formulas may as usual be carried out so that simple syntactic operations on sentences and formulas correspond to primitive recursive functions on their code numbers. I write T[A] to mean T(a), where a is the numeral for the code number of A. Otherwise I follow the relaxed attitude towards notation in Sheard's "Guide to Truth Theories" [1994].

The base theory will be first-order Peano arithmetic PA, with the understanding that when new predicates are added to the language, the instances of the scheme of mathematical induction for formulas involving them are added as well. The underlying logic will be classical, and where it makes a difference it may be assumed that the deduction system for classical logic is one in which proofs do not involve open formulas, and the only rule is modus ponens. Even in weak subtheories of PA, notions of *correctness* and *erroneousness* can be defined for atomic arithmetic sentences, which are equations between closed terms, and proved to have the properties one would expect for truth and falsehood restricted to such sentences. And even in such weak subtheories,

construction of self-referential examples is possible by the usual diagonal procedure. These include *truth-tellers*, asserting their own truth, and two kinds of *liars*, namely, *falsehood-tellers* asserting their own falsehood, and *untruth-tellers*, asserting their own untruth. (Here talk of a sentence "asserting" such-and-such really means the sentence's being provably equivalent in the theory to such-and-such.) Unlike Friedman and Sheard I will not count any truth principles — they count truth distribution and truth classicism — as part of the base theory. Comments on some individual principles will be in order.

As to the four rules, these are, like the rule of necessitation in modal logic, to be applied only in categorical demonstrations, not hypothetical deductions. For instance, with truth introduction, if we have *proved* that *A*, we may infer "*A* is true". If we have merely deduced *A* from some hypothesis, we may not infer "*A* is true" under that hypothesis. Allowing introduction or elimination to be used hypothetically would amount to adopting truth appearance and disappearance, and hence truth transparency, as axiom schemes applicable to all sentences, and that would be inconsistent. Indeed, the usual reasoning in the liar paradox shows that allowing either one of introduction or elimination to be used hypothetically, while allowing the other to be used at least categorically, leads to contradiction.

As to the axioms and schemes, the composition and decomposition axioms, even without those for atomic truth and falsehood, imply truth transparency for *arithmetical* sentences and formulas, arguing by induction on logical complexity of the sentence or formula in question. With composition and decomposition for atomic truth and falsehood as well, truth transparency extends to *truth-positive* sentences and formulas, those built up from arithmetical formulas and atomic formulas involving the new predicates by conjunction, disjunction, and quantification. With the further addition of truth consistency, one would get truth distribution and truth disappearance for all formulas.

3

The other side of axiomatic truth theory originates with Solomon Feferman. The background here is his well-known work on predicative analysis (Feferman [1964]). The idea of predicative analysis is that one starts with the natural numbers, and then considers a first round of sets of natural numbers defined by formulas involving quantification only over natural numbers, and then considers a second round of sets of natural numbers defined by formulas involving quantification only over natural numbers and sets of the first round, and so on. The process can be iterated into the transfinite, up to what has come to be called the Feferman-Schütte ordinal Γ_0.

Instead of considering round after round of sets, those of each round defined in terms of those of earlier rounds, one could consider instead round after round of satisfaction predicates, each applying only to formulas involving only earlier ones. Instead of speaking of definable sets and elementhood one would speak of defining formulas and satisfaction. But in arithmetic formulas can be coded by numbers, and the notion of the satisfaction of a formula by a number reduced to that of the truth of sentence obtained by substituting the numeral for the number for the variable in the formula. So in the end all that is really needed is round after round of truth predicates, each applicable only to sentences containing only earlier ones. Feferman [1991] finds that the process iterates only up to the ordinal ε_0, though by introducing what he calls "schematic" theories it can be extended up to Γ_0.

Kripke gives a set-theoretic construction of a model for a language with a self-applicable truth predicate, and this raises the question whether the hierarchy of truth predicates could be replaced by a single self-applicable one. To pursue this possibility it would be necessary to replace the set-theoretic construction of a model by an axiomatic theory. Thus arose the question of *axiomatizing Kripke's theory of truth*.

Feferman proposed a candidate axiomatization (which became known from citations of his work in the literature well before its publication in Feferman [1991]) with

all the composition and decomposition axioms. In the literature the label KF (for Kripke-Feferman) is sometimes used for this theory, as it will be here, but is sometimes used for this theory plus truth consistency, which here will be called KF$^+$. Later Volker Halbach and Leon Horsten [2006] produced a variant of KF based on partial logic, which they called PKF but which I will call KHH. They give a sequent-calculus formulation, but a natural deduction formulation will be given in a book by Horsten [forthcoming].

4

This past semester an undergraduate philosophy major at my school, Dylan Byron, asked me to direct him in a reading course on the literature on axiomatic theories of truth. Over the semester he expressed increasing disappointment at the scarcity in the literature of articulations of just what the philosophical aims and claims of axiomatic truth theories are supposed to be, and hearing his complaints I became convinced that there was a need for more philosophical discussion of just what is meant by "an axiomatization of Kripke's theory of truth".

There are at least three potential sources of ambiguity, two generally recognized and the other perhaps other not. To begin with, Kripke has not just one construction, but several, differing in two dimensions. On the one hand, one can choose among different underlying logical schemes: the Kleene trivalent scheme, the van Fraassen supervaluation scheme, and others. On the other hand, for any given scheme, one can choose among different fixed points: the minimal one, the intersection of all maximal ones, and others. The multiplicity of fixed-points is what allows Kripke to distinguish the outright paradoxical examples like liar sentences from merely ungrounded examples like truth-teller sentences, the former being true in no fixed points, the latter in some but not others. These two sources of ambiguity in the notion of "Kripke's theory of truth" are generally recognized. It is the minimal fixed point on the Kleene scheme that has received the most

attention, from Kripke's original paper to the present day — I set aside work of Andrea Cantini [1990] on the van Fraassen scheme — and I will concentrate on it.

Beyond this, though it would be difficult to overstate how guarded are Kripke's philosophical formulations in his "Outline", one passage does suggest that there may be two levels or stages of understanding the concept of truth, earlier and later:

> If we think of the minimal fixed point, say under the Kleene valuation, as giving a model of natural language, then the sense in which we can say, in natural language, that a Liar sentence is not true must be thought of as associated with some later stage in the development of natural language, one in which speakers reflect on the generation process leading to the minimal fixed point. It is not itself a part of that process. (Kripke [1975], 714)

Thus there is a further ambiguity in the notion of "axiomatizing Kripke's theory of truth", and a need to distinguish the problem of codifying in axioms a pre-reflective understanding of truth from the problem of doing the same for a post-reflective understanding.

5

Early in Kripke's exposition of his proposal (§III of Kripke [1975]), he invites us to join him in imagining trying to explain the meaning of "true" to someone who does not yet understand it. Herein lies what for me is a crucial question for the problem of axiomatizing the earlier, pre-reflective understanding, which I would state as follows:

> *Internal Axiomatization*. What is the simplest and most natural set of axioms and rules whose adoption by a subject who had never heard the word "true" before would give that subject an understanding of truth for which the minimal fixed point on the Kleene scheme would be a good model?

If we had an answer to this question, the question whether the minimal fixed point on the Kleene scheme really provides a good "model of natural language" would largely reduce to the question whether it is plausible to suggest that speakers of natural language first acquire an understanding of truth by adopting something like the indicated system of axioms and rules. Needless to say, the notion of "good model" here is an intuitive, not a rigorously defined one.

The internal axiomatization question is essentially the question of what we would have to tell a subject who had never heard the word "true" before to help him acquire a pre-reflective understanding of Kripkean truth. One might be inclined to think, "We could just tell *him* what Kripke tells *us*." But Kripke, as he repeatedly emphasizes, is speaking to *us* in a metalanguage, describing his fixed points from the outside, saying things that cannot be said in the object language, or recognized as true from the inside. Kripke says, for instance, that neither untruth-teller sentences nor truth-teller sentences are true, thus asserting what an untruth-teller sentence asserts and denying what a truth-teller sentence asserts. If we told the subject what Kripke tells *us*, we'd be skipping right over the pre-reflective to the post-reflective stage.

The problem of axiomatizing the later, *post*-reflective understanding, is a separate problem, which I would state as follows:

> *External Axiomatization*. What is the simplest and most natural axiomatic replacement for Kripke's set-theoretic definition of the minimal fixed point on the Kleene scheme?

The notion of "simplest and most natural axiomatic replacement" is no more rigorously defined than that of "good model", but this does not mean that we cannot recognize examples when we see them. A paradigm would be PA itself, arguably the simplest and most natural set axiomatic replacement for the set-theoretic definition of the natural numbers as the elements of the smallest set containing zero and closed under successor.

6

Beginning with the internal question, let us return to Kripke's discussion of the subject being taught the meaning of "true" (Kripke [1975], 701). Kripke supposes the subject has knowledge of various empirical facts: for instance, meteorological facts, such the fact that snow is white, and historical facts about what is said in what texts, perhaps the fact that "Snow is white" appeared in the *New York Times* on such-and-such a date. But the subject has initially no knowledge about truth. Kripke then imagines us telling the subject "that we are entitled to assert (or deny) of any sentence that it is true precisely under the circumstances when we can assert (or deny) the sentence itself", which I take to amount to giving him the four categorical rules of inference in the table.

Kripke then explains how his subject, having already been in a position to assert "Snow is white", is now in a position to assert "'Snow is white' is true", and how, having already been also in a position to assert "'Snow is white' appears in the *New York Times* of such-and-such a date", he is now in a position to infer and assert "Some true sentence appears in the *New York Times* of such-and-such a date". Kripke concludes "In this manner, the subject will eventually be able to attribute truth to more and more statements involving the notion of truth itself."

Kripke's discussion can be adapted to the situation where the base theory to which the truth predicate is being added is PA. We suppose the subject initially knows and speaks of nothing but numbers and their arithmetical properties, and of sentences and their syntactic properties insofar as statements about the latter can be coded as statements about the former. Now suppose we introduce a truth predicate and give the subject the four categorical rules in the table. Let us call the resulting theory PA*.

Then what Kripke said about "Snow is white" and "…appears in the *New York Times* of such-and-such a date" applies to, say, "Seventeen is prime" and "…is provable in Robinson arithmetic Q". The subject will be able to assert — the theory PA* will be

able to prove — that Robinson arithmetic proves some true sentence, and beyond that "more and more statements involving the notion of truth itself".

7

The paper of Friedman and Sheard contains information about the scope and limits of what PA* can prove. In the first place, it can't prove contradictions: it is consistent, as is the theory, now called FS (for Friedman-Sheard), which adds truth consistency and completeness, and the composition and decomposition axioms except those for atomic truth and falsehood. Consistency is proved by a model-theoretic construction that represents an independent discovery of the principle of "revision" theories of truth.

To recall how revision works in a fairly general form (as in various works of Anil Gupta and Hans Herzberger), we can construct a sequence of models, indexed by ordinals, each of which consists of the standard model of arithmetic plus an assignment of an extension to the truth predicate. At stage zero the truth predicate may be assigned any extension we please. At stage one its extension consists of the (code numbers for) sentences that are true in Tarski's sense at stage zero. Stage two is obtained from stage one as stage one was from stage zero, and so on. At stage ω, anything that has always been true from some point on is put in the extension, and anything that has always been false from some point on is left out of the extension. Other sentences are put in or left out according as they were put in or left out originally, at stage zero. And so on.

The consistency proof in the paper of Friedman and Sheard involves only the finite stages. One considers the set that contains a sentence *A* just in case *A* has always been true from some point on. This set is closed under logical consequence and under the four categorical truth rules, and contains all the axioms and theorems of PA* and indeed of FS, but does not contain 0 = 1. In the section of the paper on proof-theoretic strength, a refinement of the method of the consistency proof is used to show that PA* is a

conservative extension of PA. It proves no new arithmetical sentences. (Indeed, this is proved for PA* plus the axioms of truth consistency and truth completeness.)

The revision method can be adapted to show that PA* by itself does not imply various additional axioms of FS. For instance, let *L* be a liar sentence. Start at stage zero with *L* in the extension of the truth predicate and its conjunction with itself, *L* & *L*, out of it. Continue the construction through all the ordinals less than ω^2. Consider the set of sentences that have always been true from some point on. These will not contain the sentence "if *L* is true then *L* & *L* is true", because this will fail at stage zero and at every subsequent limit stage, and positive conjunctive composition therefore fails. Something similar can be done for truth consistency and completeness and the composition and decomposition axioms for the connectives and quantifiers.

8

PA*, though suggested by a literalistic reading of some of Kripke's initial heuristic remarks, does not correspond very directly to Kripke's eventual set-theoretic construction of the minimal fixed point on the Kleene scheme, and this for a double reason.

First, PA*, like all the systems considered by Friedman and Sheard, is based on classical logic, and has every instance of truth classicism as a theorem, including for example $\mathsf{T}[L \vee \sim L]$ where *L* is a liar sentence. Kripke represents himself as adhering throughout to classical *propositional* logic, but allowing that departures from classical *sentential* logic may be needed if our language contains sentences that do not express propositions (as departures from classical sentential logic would also be needed if our language contained ambiguous sentences expressing multiple propositions). Even where there are sentences that do not express propositions, use of classical logic will be appropriate if one follows the van Fraassen scheme; but on the Kleene scheme, $\mathsf{T}[L \vee \sim L]$, where *L* is a liar sentence, does *not* hold in any fixed point, and the internal axiomatization question as I formulated it was a question about the Kleene scheme.

Presumably this first problem could be resolved by replacing PA* with a version pPA* based on partial logic. And in any case it is of interest to consider the internal axiomatization question for the van Fraassen scheme.

Second, however, even considering the question for the van Fraassen scheme, the minimal fixed point does not provide a good model for PA*. Though as acknowledged earlier, the notion of "good model" is not a rigorously defined one, that does not prevent us from recognizing that the minimal fixed isn't one, simply because it is far more complicated than is needed. We get a model of PA* already even if we only carry out the finite stages of Kripke's inductive construction. Even the simplest and most natural examples of sentences that don't get evaluated as true or false until some transfinite stages turn out to be neither provable nor refutable PA*. For example, if we let τ_0, τ_1, τ_2, … be the sentences $0 = 0$, $T[0 = 0]$, $T[T[0 = 0]]$, … and let τ_ω be the sentence saying that all the τ_n are true, then PA* cannot prove τ_ω. (It fails in the model used to prove consistency.)

Thus we have not found in the considerations advanced so far an answer to the internal axiomatization question, the question what to tell the subject who does not know the meaning of "true". Telling him everything Kripke tells us is too much, and telling him just the four categorical rules is not enough.

9

At this point, a fantasy suggests itself. Suppose that instead of starting with a human being and giving *him* the four categorical truth rules, we start with a Superhuman Being, and give *Her* those rules. We suppose She has enormous cognitive abilities about all matters *not* involving the notion of truth, which in the test case of arithmetic might be represented by the ability to draw inferences using the omega rule.

For any arithmetic sentence *A* that is true in Tarski's sense, She can prove it using the ω-rule, and She can then infer "*A* is true". If *A* is unprovable in PA, then the

arithmetical sentence saying so will be true in Tarski's sense, so She can prove that, too, using the ω-rule. So She can prove "Some true arithmetical statement is not provable in PA", and so on.

Formally we might represent Her by a theory PAω* consisting of Peano arithmetic plus the omega rule and the four categorical truth rules. It is not too hard to see (using the basic result about ω-logic that a sentence follows by ω-logic from a set of first-order sentences if and only if it is true in all ω-models of that set) that what is provable is precisely what holds in the minimal fixed point on the van Fraassen scheme.

Presumably by replacing our theory with a variant pPAω* based on partial logic we can get an equivalent characterization of the minimal fixed point on the Kleene scheme. But needless to say, none of this gives us an answer to the internal axiomatization question as I formulated it, as a question about natural language as spoken by human beings, not Superhuman Beings.

10

Another thought may now suggest itself. Perhaps we could tell our human subject about the foregoing fantasy, and then in addition specify that he is entitled to assert a sentence himself if and only if he is entitled to assert that She of the fantasy would be entitled to assert it. Formally, we could add a predicate S for "The Superhuman Subject could assert", with appropriate axioms and rules.

The question which principles are appropriate for S must be approached with caution, however. We cannot, for instance, assume "If She can assert that *A*, then *A*" as an unrestricted axiom scheme, since contradiction results upon applying that principle to a self-referential sentence of the kind "This very sentence is something She cannot assert". The most cautious approach would assume that S[*A*] or T[*A*] hold only for sentences *A* not containing S (as with a Tarski-style truth-predicate).

Some principles that seem appropriate are the following: (a) the rules permitting inference in categorical demonstrations from "She can assert that *A*" to *A* and from *A* to "She can assert that *A*", as per our imagined specifications to the human subject; (b) the axiom that She can assert any axiom of logic or arithmetic; (c) the axiom that She can make inferences from assertion to assertion using modus ponens; (d) ditto for the ω-rule; (e) ditto for the four truth rules. Let us call the system given by these principles SPAω*.

SPAω* is consistent. This can be established by showing that a fixed point on the van Fraassen scheme provides a model (à la van Fraassen). The arithmetical part of the model is standard. The predicates T and S have the same extension, the set of (codes for) sentences valued true in the fixed point, but T is treated as a partial predicate, with anti-extension the set of (codes for) sentences valued false, whereas S is treated as a total predicate, whose anti-extension is simply the complement of its extension.

SPAω* provides more than enough in the way of axioms and rules to prove the test sentence τ_ω mentioned earlier as unprovable in PA*. We may reason as follows. We can assert $0 = 0$ or τ_0, hence so can She. But then since She can reason using the truth rules, for any n, if She can insert τ_n, then She can assert $\mathsf{T}[\tau_n]$ or τ_{n+1}. Hence, by induction, for every n, She can assert τ_n, and that τ_n is true. Hence, since She can reason by the ω-rule, She can assert that for every n, τ_n is true. But that is to assert τ_ω, and since we have just deduced that She can assert it, we can assert it, too.

SPAω* provides enough in the way of axioms and rules to prove the consistency of PA as well. The argument is that She can assert each axiom, and She can reason by modus ponens so She can assert each theorem, and so through Her ability to use the truth rules, She can assert *the truth of* each theorem of PA, and since — skipping some details here — She can also assert for each nontheorem that it is *not* a theorem, She then can, for each sentence, assert that if it is a theorem it is true, and then through Her ability to use the ω-rule, She can assert that every theorem of PA is true. But $0 \neq 1$, and since we have just asserted it, She can assert that $0 \neq 1$ as well, and then through Her ability to use the

truth rules, She can infer that 0 = 1 is untrue. Hence She can assert that 0 = 1 is a nontheorem, and since we have just deduced that She can assert it, we can assert it, too.

I will not pursue the development of the theory further here. In particular, I leave the determination of the exact proof-theoretic strength of SPAω*, and that of the variant pSPAω* based on partial logic, to the experts. Presumably pSPAω* would represent one candidate answer to what I have called the internal question, the question of what to tell the human subject who has never heard the word "true" before. But having brought this kind of answer, involving a new predicate over and above the truth predicate, to your attention, let me now set it aside.

11

Returning to theories involving no new predicates but T, there lie near at hand two further conceivable answers to the question what to tell our subject: "We can tell him the axioms of KF" or "We can tell him the axioms of KHH". (I do not mean to imply that the originators of either theory advocated it as an answer to the internal question as I have posed it, but only that it is natural to take up the issue whether one or the other of them might be a good answer.)

The difference between the two is that KF is based on classical, and KHH on partial logic. This difference results in a difference in proof theoretic strength. For Halbach and Horsten [2006] show that their system, though stronger than FS, which Halbach [1994] had shown to have the same strength as RA($<\omega$), is weaker than KF, which Feferman [1992] had shown to have the same strength as RA($<\varepsilon_0$). Its strength is that of RA($<\omega^\omega$).

But though it is the issue of partial *versus* classical logic that distinguishes PA* or PAω* from pPA* or pPAω*, and the same issue that distinguishes KF from KHH, the issue apparently distinguishes in different ways in the two cases. The difference between PAω* and pPAω* looks like the difference between an object language based on a van

Fraassen fixed point and an object language based on a Kleene fixed point. The difference between KF and KHH looks like the difference between an object language based on a Kleene logic and a metalanguage based on classical logic.

The mere fact that it is based on classical rather than partial logic means that KF has theorems, such as instances of excluded middle for liar sentences, that are not valued true in the minimal or indeed any fixed point on the Kleene scheme. For that matter, many of the composition and decomposition axioms of KF are not valued true in any fixed point on *either* the Kleene *or* the van Fraassen scheme. Moreover KF proves sentences such as "a liar sentence is not true" that Kripke explicitly classifies as post-reflective rather than pre-reflective. All this seems to disqualify KF as an answer to the internal axiomatization question, leaving KHH as the only surviving candidate.

If we accept KHH as tentative answer to the internal question, we now face the question whether it is plausible to suggest that the way we actually learn the meaning of "true" is by coming to internalize something like that system of axioms and rules. Perhaps the sheer number of axioms and rules involved is enough to make the defensibility of the claim seem doubtful, but in addition there is fact that, in Feferman's quotable phrase, "nothing like sustained ordinary reasoning can be carried out" in the kind of partial logic on which KHH is based. It was this fact that led Feferman seek a system like KF based on classical logic in the first place. Perhaps the availability of the natural deduction formulation in Horsten [forthcoming] may soften this judgment, but this is not an issue I can pursue further here.

12

I turn instead from the internal to the external question, the question of characterizing by axioms and rules, without explicit set-theoretic apparatus, Kripke's model construction, as viewed from the outside rather than the inside. I should first consider whether KF, which was disqualified as an answer to the internal question, looks

any better as an answer to external question, since after all the reason KF was disqualified as an answer to the internal question was precisely that it seemed to be describing a fixed point from without rather than from within.

But KF does not look good as answer to the external axiomatization question, either, at least not as I formulated the question. KF provides a simple and natural axiomatization, from an external perspective, of the properties of an arbitrary fixed point; but as I formulated the axiomatization question, it was not about *an arbitrary* fixed point but rather about *the minimal* fixed point. KF shares with KHH the feature that, by design, we get a model for it from *any* Kleene fixed point. This feature seems less appropriate for a representation of a *post*-reflective understanding, than for a representation of a *pre*-reflective understanding, when we are just going forward with whatever specifications we have been given, without thinking about the scope and limits of how far they will take us.

This feature is what is responsible for the limited proof-theoretic strength of KF. Earlier I mentioned that Feferman showed its strength to be the same as that of the theory known as $RA(<\varepsilon_0)$. But that is known (using results in Feferman [1982] in one direction and a combination of Aczel [1980] with Friedman [1970] in the other) to be the same as the proof-theoretic strength of the theory asserting for each positive arithmetic inductive operator the existence of *some* fixed point; and the more illuminating comparison is between KF and this last theory. KF is interpretable in it because the inductive operator involved in Kripke's construction is positive arithmetic, and because *any* fixed point for that operator provides a model of KF.

Interpretability holds in the opposite direction also (not literally in KF itself, but in the conservative extension related to KF as the theory known as ACA_0 is related to PA), because, as noted in Cantini [1989], for any positive arithmetic inductive operator, KF proves that a certain associated self-referential formula defines a fixed point for it. This may be shown as follows. Let $\Phi(X, x)$ be a formula with no quantification over set-

variables in which the all appearances of the free set-variable X are positive. We may take it to be of the form

(1) $\forall y_1 \ \ldots \exists y_k \bigvee_{i = 1 \text{ to } r} \bigwedge_{j = 1 \text{ to } s} \varphi_{ij}(x, y_1, \ \ldots \ , y_k)$

where each φ_{ij} is either arithmetical or of the form $X(u_{ij})$ where u_{ij} is one of the variables $y_1, \ldots , y_k$. In the usual way introduce $B(x)$ that "says"

(2) $\forall y_1 \ \ldots \exists y_k \bigvee_i \bigwedge_j \ \varphi_{ij}^*(x, y_1, \ \ldots \ , y_k)$

where φ_{ij}^* is φ_{ij} if φ_{ij} is arithmetical, and $\mathsf{T}[B(u_{ij})]$ if φ_{ij} is $X(u_{ij})$. In other words, $B(x)$ "says" $\Phi(\mathsf{T}[B()], x)$. Then KF proves

(3) $\forall x \ (B(x) \leftrightarrow \forall y_1 \ \ldots \exists y_k \ \bigvee_i \bigwedge_j \varphi_{ij}^*(x, y_1, \ \ldots \ , y_k))$

and since the formula involved is T-positive, by transparency KF proves

(4) $\forall x \ (\mathsf{T}[B(x)] \leftrightarrow \forall y_1 \ \ldots \exists y_k \ \bigvee_i \bigwedge_j \ \varphi_{ij}^*(x, y_1, \ \ldots \ , y_k))$

In other words, if we give the formula $\mathsf{T}[B(x)]$ the name $\Phi^*(x)$, then KF proves for pertinent Φ that the associated Φ^* gives a fixed point for the operator given by Φ.

13

The only attempt in the literature known to me to frame an axiomatic theory of truth that would incorporate minimality, as KF and KHH do not, occurs in work of Cantini [1989]. He introduces a system KF$^+$ + GID, where GID is a certain general scheme of inductive definition that he shows to hold for the minimal fixed point. The incorporation of minimality through such a scheme is somewhat indirect, and to that extent complicated and artificial, so I would like to propose another system, which I will call KFμ (with μ for minimal), that incorporates minimality more straightforwardly.

Let us think of our theories as formulated with only the truth predicate T as primitive (the falsehood predicate F being treated as a defined). KF takes the composition and decomposition laws in the table as axioms. By contrast, KFμ takes only the composition laws as axioms, but adds an axiom scheme of minimality. For each formula $\tau(x)$ it is an axiom that if the set of truths satisfying τ is closed under the composition laws, then every truth satisfies τ. What the closure assumption here amounts to is a conjunction of conditions, one for each composition axiom. Thus corresponding to positive equational composition, for instance, we have the condition that any correct equation satisfies τ. KFμ provides one *obvious* candidate for an answer to the external axiomatization question, and I will devote the remainder of this talk to exploring its consequences.

The first thing to be noted is that every theorem of KF and indeed of KF^+ is a theorem of KFμ. First, to get KF, we note that the decomposition axioms can be derived using the minimality scheme. Taking the positive equational decomposition axiom for example, let $\tau(x)$ say: if x is (the code number for) and equation, it is correct. It is easily seen that the set of truths satisfying τ is closed under the composition laws. The only such law that could conceivably be a problem is the one whose consequent mentions the specific kind of sentence mentioned in τ (namely, equations), since all other sentences *vacuously* satisfy τ. So we need only check the positive equational composition axiom; and the condition corresponding to this axiom is (as noted above) simply that every correct equation should satisfy τ, which it trivially does. The minimality principle now applies to tell us that every truth satisfies τ, which is to say, that every true sentence fulfills the condition that if it is an equation it is correct, or more simply, that every true equation is correct. That is the positive equational decomposition axiom we wanted. Exactly the same method can be used with all the other decomposition axioms.

Further, to get KF^+, the same method can be used to prove the truth consistency axiom, that no sentence is both true and false (where falsehood is truth of the negation).

Let $\tau(x)$ say: if x is not (the code number for) of a false sentence. It is then not hard to prove in KFμ closure under the composition laws. For instance the composition law for addition, for instance, what we must check is that if A and B are true and not false, then their conjunction is true and not false. But this is easy using the positive composition and negative decomposition axioms for conjunction (which respectively tell us that if both conjuncts are true, their conjunction is true, and that if a conjunction is false, one of its conjuncts is false). Once we have all the composition axioms, minimality applies to tell us that every true sentence satisfies the condition τ of not being false, which is the truth consistency axiom. That axiom is known to be unprovable in KF, though it has also been shown by Cantini that KF$^+$ is not of greater proof-theoretic strength than plain KF.

14

There are also theorems of KFμ that are not theorems of KF$^+$. Consider a truth-teller sentence, a sentence B constructed by the usual diagonal method so that B "says" $\mathsf{T}[B]$. It cannot be proved in KF$^+$ that B is not true, because the axioms of KF$^+$ hold for any fixed point, and though no truth-teller is true in the *minimal* fixed point, each is true in some fixed point or other. But it can be proved in KFμ that B is not true.

What B literally *is*, is a sentence $\forall z(\sim C(z) \vee \mathsf{T}(z))$ with code b, where C is arithmetical and where $C(b)$ and $\forall n(n = b \vee \sim C(n))$ are provable in PA, so that PA proves $\mathsf{T}(b) \rightarrow B$ and $B \rightarrow \mathsf{T}(b)$. Note that these are provable in KFμ not only because KFμ extends PA, but also simply because B is T-positive. Now let $\tau(x)$ say: x is not (the code number for) any of the following formulas:

$\mathsf{T}[B] = \mathsf{T}(b)$

$\sim C(b) \vee \mathsf{T}(b)$

$B = \forall p(\sim C(p) \vee \mathsf{T}(p))$

Let $\mathsf{T}^*(x)$ abbreviate the conjunction of $\mathsf{T}(x)$ and $\tau(x)$. Checking the closure of the set of sentences satisfying τ under the composition laws amounts to checking that the composition axioms hold with truth* (or T^*) in place of truth (or T). The only cases that could possibly cause trouble are those where the consequent of the axiom is the truth* of one of the three exceptional sentences mentioned in τ:

(1) if B is true*, then $\mathsf{T}[B]$ is true*

(2) if $\sim C(b)$ is true* or $\mathsf{T}(b)$ is true*, then $\sim C(b) \vee \mathsf{T}(b)$ is true*

(3) if for all p, $\sim C(p) \vee \mathsf{T}(p)$ is true*, then $\forall p(\sim C(p) \vee \mathsf{T}(p))$ is true*

In all these cases the consequent of the axiom fails by definition of T^*, so we must show, working in KFμ, that the antecedent fails as well. And indeed, the antecedent of (1) fails by definition of truth*. The first disjunct of the antecedent of (2) fails since $C(b)$ holds and we have truth transparency for $\sim C(b)$ because it is arithmetical, while the second disjunct fails by definition of truth*. The antecedent of (3) fails in the instance $p = b$ by definition of T^*. Once we have the all the composition axioms, minimality applies and tells us that every true sentence is true*, which is to say, is not one of the sentences mentioned in T^*, or in other words, that those sentences, including B, are not true. In a similar way it can be proved that B is not false, either.

15

KFμ is essentially a subtheory of the theory known in the literature as ID_1, which for each positive arithmetic inductive operator asserts the existence of a minimal fixed point. In the joint paper with Sheard, Friedman proves that a certain axiomatic truth theory H is proof-theoretically as strong as ID_1.

(It may be mentioned that though KF$^+$ has all the truth principles from the list of Friedman and Sheard for the system they call H, as well as truth distribution, which is part of the Friedman-Sheard base theory, and though Friedman proves that H is of

impredicative proof-theoretic strength, and so of much greater strength than KF, still Friedman's proof for H does not apply to KF$^+$, because it makes use of truth classicism, which is part of the Friedman-Sheard base theory, and so not mentioned by them explicitly in their definition of H, but is not available in KF$^+$.)

Friedman's proof uses truth principles only to get a certain lemma, after which they are not referred to again. So it will be enough for us to prove for T = KFμ the lemma that Friedman proves for T = H. It reads as follows:

> *Friedman's Lemma*. For any arithmetical formula $R(x, y)$ and any formula $A(x)$, the theory T proves the following:
> If transfinite induction along R holds for $\mathsf{T}[B(x)]$ for every formula $B(x)$
> then transfinite induction along R holds for $A(x)$.

The claim that transfinite induction holds for all formulas of form $\mathsf{T}[B(x)]$, unlike the claim that transfinite induction holds for *all* formulas $A(x)$, can be stated in a single sentence, and that is what makes the lemma key to Friedman's proof.

There is in fact a single formula B, depending only on R, such that for any formula A, KFμ proves that if transfinite induction along R holds for $\mathsf{T}[B(x)]$, then in holds for $A(x)$. The formula is suggested by Kripke's proof that the set of truths in the minimal fixed point is a complete Π^1_1 set.

Take B constructed by the usual diagonal method so that $B(n)$ "says" $\forall m(R(m, n) \rightarrow \mathsf{T}[B(m)])$. What $B(x)$ literally *is*, is a formula

$$\forall z \forall y \forall w(\sim C(z) \vee \sim R(y, x) \vee \sim I(z, y, w) \vee \mathsf{T}(w))$$

with code b, where C is arithmetical, where $C(b)$ and $\forall n(n = b \vee \sim C(n))$ are provable in PA, and where I is the usual arithmetical formula expressing "w is the code for the sentence resulting from substituting the numeral for y for the free variable in the formula with code z", with its usual properties. Then PA proves

$$\forall n(\forall m(R(m, n) \rightarrow \mathsf{T}[B(m)]) \rightarrow B(n))$$

and its converse. Note that KFμ then proves

$$\forall n(\forall m(R(m, n) \rightarrow \mathsf{T}[B(m)]) \rightarrow \mathsf{T}[B(n)])$$

by truth transparency, since B is T-positive. This is the antecedent of transfinite induction along R for $\mathsf{T}[B(x)]$, so by the assumption of Friedman's lemma the consequent $\forall n\mathsf{T}[B(n)]$ holds.

Suppose now for contradiction that the conclusion of Friedman's lemma fails, that the antecedent of transfinite induction along R for $A(x)$, namely,

$$\forall n(\forall m(R(m, n) \rightarrow A(m)) \rightarrow A(n))$$

holds, but the consequent $\forall nA(n)$ fails. To deduce a contradiction and complete the proof that Friedman's lemma holds, we prove, working in KFμ, that for any n, $\sim A(n)$ implies $\sim\mathsf{T}[B(n)]$ or equivalently $\sim B(n)$.

To this end, writing b_n for the code of $B(n)$,l et $\mathsf{T}^*(x)$ be the conjunction of $\mathsf{T}(x)$ and the formula $\tau(x)$ saying that x is not (the code of) any of the following sentences:

$\mathsf{T}[B(n)] = \mathsf{T}(b_n)$, where $\sim A(n)$
$\sim C(b) \vee \sim R(m, n) \vee \sim I(b, m, b_m) \vee \mathsf{T}(b_m)$, where $R(m, n)$ and $\sim A(m)$
$B(n) = \forall z\forall y\forall w(\sim C(z) \vee \sim R(y, x) \vee \sim I(z, y, w) \vee \mathsf{T}(w))$, where $\sim A(n)$

In verifying the composition axioms for T^*, the only cases that could conceivably cause trouble are those where the consequent of the axiom is the truth of one of the three sentences mentioned in T^*:

(1) if $B(n)$ is true*, then $\mathsf{T}[B(n)]$ is true*

(2) if $\sim C(b) \vee \sim I(b, m, b_m) \vee \sim R(m, n)$ is true* or $\mathsf{T}(b_m)$ is true*
then $\sim C(b) \vee \sim R(m, n) \vee \sim I(b, m, b_m) \vee \mathsf{T}(b_m)$ is true*

(3) if for all p, m, q, $\sim C(p) \vee \sim R(m, n) \vee \sim I(p, m, q) \vee \mathsf{T}(q)$ is true*
then $B(n) = \forall p \forall m \forall q(\sim C(p) \vee \sim R(m, n) \vee \sim I(p, m, q) \vee \mathsf{T}(q))$ is true*

Here (1) will be troublesome when $\sim A(n)$, (2) will be troublesome when $R(m, n)$ and $\sim A(m)$, and (3) will be troublesome when $\sim A(n)$ again, since these are the cases in which the consequent of the axiom fails by the definition of T* and we must prove in these cases that the antecedent fails also. For (3), if $\sim A(n)$, then by the antecedent of transfinite induction along R for $A(x)$, there is some m with $R(m, n)$ and $\sim A(m)$, and the thing to prove is that the antecedent of (3) fails for $p = b$ and this m and $q = b_m$. The argument is much as in the case of the truth-teller example.

Once we have the all the composition axioms, minimality applies and tells us that every true sentence is true*, which is to say, is not one of the sentences mentioned in T*, or in other words, that those sentences, including $B(n)$ whenever $\sim A(n)$, are not true. So Friedman's lemma is established, and applies to complete the proof.

To show that ID_1 is outright interpretable in (a conservative extension of) KFμ, it would be enough to prove a metatheorem to the effect that for any pertinent formula Φ there is a formula Φ^* such that it can be proved in KFμ that Φ^* gives a minimal fixed point for the operator given by Φ. A formula Φ^* was introduced towards the end of §12 above, and shown to give a fixed point. It would only remain to show, applying the minimality principle of KFμ as in this section and the preceeding, that this fixed point is *minimal*. But this is not the place to go into details.

16

Having taken so much from Friedman, it may be bad form for me to close by asking for more, but I am tempted to do so. It would be nice to have the Friedman-

Sheard project redone in a way that does not make truth classicism or any truth principles as part of the base theory. It would be nice to have the Friedman-Sheard project done again in a version based on partial rather than classical logic. I would be nice to have the Friedman-Sheard project redone for theories with *two* predicates T and S. The task is enormous, and so one may think, "If only Friedman could be lured back to the subject to do some of this work for us!"

But perhaps philosophers should first do a bit more in the way of distinguishing, within the enormous range of combinations, those potentially of the most philosophical interest. Something like that is what I have been attempting to do here.

REFERENCES

Aczel, Peter, 1980, "Frege structures and the notion of proposition, truth and set", The Kleene Symposium, Jon Barwise et al. (editors), Amsterdam: North-Holland, 31-59.

Cantini, Andrea, 1989,"Notes on Formal Theories of Truth", *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 35:97-130.

———————, 1990, "A Theory of Formal Truth Arithmetically Equivalent to ID_1", *Journal of Symbolic Logic* 55:244-59.

Feferman, Solomon, 1964, "Systems of Predicative Analysis", *Journal of Symbolic Logic* 29:1-30.

———————, 1982, "Iterated Inductive Fixed-Point Theories: Application to Hancock's Conjecture", in G. Metakides (ed.), *Patras Logic Symposium*, Amsterdam: North Holland, 171-196.

———————, 1991,"Reflecting on Incompleteness", *Journal of Symbolic Logic* 56:1-49.

Friedman, Harvey, 1970, "Iterated Inductive Definitions and $\sum^1_1$-AC", in J. Myhill (ed.), *Intuitionism and Proof Theory*, Amsterdam: North Holland, 435-442.

Friedman, Harvey and Michael Sheard, 1987, "An Axiomatic Approach to Self-Referential Truth", *Annals of Pure and Applied Logic* 33:1-21.

———————, 1988, "The Disjunction and Existence Properties for Axiomatic Systems of Truth", *Annals of Pure and Applied Logic* 40:1-10.

Halbach, Volker, 1994, "A System of Complete and Consistent Truth", *Notre Dame Journal of Formal Logic* 35:311-27.

Halbach, Volker and Leon Horsten, 2006 "Axiomatizing Kripke's Theory of Truth", *Journal of Symbolic Logic* 71: 677-712.

Horsten, Leon, forthcoming, *Axiomatic Truth Theories and Deflationism*.

Leigh, Graham, and Michael Rathjen, forthcoming, "An Ordinal Analysis of Self-Referential Truth".

Sheard, Michael, 1994, “A Guide to truth Predicates in the Modern Era”, *Journal of Symbolic Logic* 59:1032-54.

PRINCIPLES OF TRUTH

RULES	
Truth Introduction	from a sentence to infer the truth of the sentence
Truth Elimination	from the truth of a sentence to infer the sentence
Untruth Introduction	from the negation of a sentence to infer the untruth of the sentence
Untruth Elimination	from the untruth of a sentence to infer the negation of the sentence
AXIOMS	
Truth Consistency	No sentence is both truth and false.
Truth Completeness	Every sentence is either true or false.
Positive Composition	
Equational	If an equation is correct, it is true.
Negational	If a sentence is false, its negation is true.
Conjunctive	If two sentences are true, their conjunction is true.
Disjunctive	*analogous*
Universalizing	*analogous*
Existentializing	*analogous*
Atomic Truth	If a sentence is true, it is true that it is true
Atomic Falsehood	If a sentence if false, it is true that it is false
Negative Composition	*correlates for falsehood of Positive Composition*
e.g. Conjunctive	If either of two sentences is false, their conjunction is false.
Positive Decomposition	*converses of Positive Composition*
e.g. Conjunctive	If a conjunction of two sentences is true, both of them are true.
Negative Decomposition	*converses of Negative Composition*
e.g. Conjunctive	If a conjunction of two sentences if false, at least one of them is false.
Truth Distribution	If a conditional and its antecedent are true, so is the consequent.
Truth Classicism	Any instance of excluded middle (or other classical law) is true.
SCHEMES	
Truth Transparency	
Sentential	A is true iff A
Formulaic	For all n, $A(n)$ is true iff $A(n)$
Truth Appearance	
Sentential	If A, then A is true.
Formulaic	For all n, if $A(n)$, then $A(n)$ is true
Truth Disappearance	
Sentential	If A is true, then A.
Formulaic	For all n, if $A(n)$ is true, then $A(n)$

THEORIES		STRENGTH
PA*	truth rules	PA
FS	truth rules, consistency & completeness, composition & decomposition except atomic truth & falsehood	$RA(<\omega)$
KF	all composition and decomposition	$RA(<\varepsilon_0)$
KF^+	KF + consistency	$RA(<\varepsilon_0)$
KHH = PKF	partial logic variant of KF	$RA(<\omega^\omega)$
KFμ	all composition + minimality	ID_1