THE REVIEW OF SYMBOLIC LOGIC
Volume 5, Number 4, December 2012

REVISION REVISITED

LEON HORSTEN
University of Bristol

GRAHAM E. LEIGH
University of Oxford

HANNES LEITGEB
Ludwig Maximilians-Universität München
and

PHILIP WELCH
University of Bristol

Abstract. This article explores ways in which the Revision Theory of Truth can be expressed in the object language. In particular, we investigate the extent to which semantic deficiency, stable truth, and nearly stable truth can be so expressed, and we study different axiomatic systems for the Revision Theory of Truth.

§1. New questions for the Revision Theory of Truth. The Revision Theory of Truth is a class of models for the language of truth ($\mathcal{L}_T$). This language of truth is intended to be a toy model for a natural language such as English. It is intended to contain all the features that are relevant for the logical properties of the notion of truth, and no more than that. $\mathcal{L}_T$ contains the first-order language of arithmetic ($\mathcal{L}_{PA}$), so as to enable sentences to talk about themselves relative to some coding scheme. In addition, $\mathcal{L}_T$ contains a truth predicate T. This predicate is intended to be a truth predicate not only for the language of arithmetic, but for the whole of $\mathcal{L}_T$.

The revision theory defines formal notions of truth, falsehood, and semantic deficiency ("paradoxicality") for $\mathcal{L}_T$. These notions are intended to be the analogues for $\mathcal{L}_T$ of the notions of truth, falsehood, and semantic deficiency of natural languages.

The basic facts about the revision theory are described in Gupta & Belnap (1993), which has become the *locus classicus* on this subject. Detailed information about the complexity of the revision theoretic notions of truth is given in Welch (2001).

The aim of this article is to contribute to the deeper analysis of the revision theory in the light of recent developments in the field of theories of truth. This analysis is inspired by an overall standpoint that is different from that of Gupta & Belnap (1993).

Since the revision theory is a class of models, it belongs, like Kripke (1975), to the category of semantic theories of truth. The revision theory is defined in a richer metalanguage: the language of set theory. For familiar Tarskian reasons this is essentially so: the revision

Received: July 26, 2011.

doi:10.1017/S175502031100030X

theory for $\mathcal{L}_T$ cannot be defined in $\mathcal{L}_T$ itself. A standard complaint against semantic truth theories is that they are expressed in an "essentially richer metalanguage." If we want to construct a truth theory for English, so the argument goes, this theory has to be formulated in English: we cannot jump outside our language. So one of the constraints on the construction of a truth theory of our toy language $\mathcal{L}_T$ is that it should be formulated in $\mathcal{L}_T$ itself.

Gupta and Belnap agree with Kripke that the notions of semantic deficiency, truth, and falsehood for $\mathcal{L}_T$ "correspond to a later stage in the development of the language" (Kripke, 1975). This amounts to the position that, while a language user can use the concept of truth to make correct statements about her own language, it exceeds her powers to formulate a correct truth theory for her own language.

Today, many authors regard the Kripkean point of view as unsatisfactory.[1] In their view, the models for $\mathcal{L}_T$ that are defined in the metalanguage are secondary. They argue that, in order not to fall prey to revenge problems, the final theory should be expressed in the object language. This entails a shift in attention to the notions that can be expressed and the theorems that can be proved in $\mathcal{L}_T$. Much of the recent work in truth theories is more in agreement with this view (*cf.* Field, 2008) than with the Kripkean view.

The new perspective can be adopted to semantic theories that were originally proposed from the Kripkean point of view. For Kripke's own theory of truth, this was done, in different ways, in Feferman (1991) and in Halbach & Horsten (2005). In this article, we want to do the same for the Revision Theory of Truth.

One of the questions that Field (2008) has brought to the foreground is:

Question 1 *To what extent can a given theory of truth expressing the notion of semantic deficiency be expressed in the object language?*

This is a question that is just as relevant for the Revision Theory of Truth as for Field's own theory of truth. In Welch (201?), the situation for Field's theory of truth is investigated. The main theme of the first part of the present article is that, from a semantic point of view, the Revision Theory is very similar to Field's theory of truth. Field argues that semantic deficiency is an irrevocably fragmented notion. He has developed a hierarchy of deficiency predicates (Field, 2008). Welch (201?) shows that the maximal length of this hierarchy is exactly the first recurring ordinal ζ of the revision sequence of models. Moreover, there are "super-liar" sentences (there called "ineffable liar" sentences) which escape the Fieldian hierarchy of deficiency predicates altogether. In the first part of this article it is shown that the situation for the revision theory is exactly the same. And this means that the revision theory, like Field's truth theory, is unable to fully capture semantic deficiency even with a hierarchy of object language concepts.

A second question that emerges from the new perspective is:

Question 2 *To what extent does the revision theory give rise to attractive axiomatic truth theories formulated in the object language?*

Attempts have been made to capture the spirit of versions of the semantic truth theory in Kripke (1975) in axiomatic theories formulated in the object language (Feferman, 1991, 2008; Halbach & Horsten, 2006). Of course Kripke's semantic account cannot be captured

[1] This critique of the Kripkean position is discussed in Horsten (2011, Chapter 2) and in Halbach (2011, Part I). For the debate between the different points of view, see the essays in Beall (2007).

completely. For one thing, all Kripke's fixed point models are based on the standard natural number structure; thus this class is not recursively axiomatizable. Nonetheless, Feferman's axiomatization *KF* (for "Kripke–Feferman") captures Kripke's account in a weaker sense. Every model of *KF* that is based on the natural numbers is a fixed point of Kripke's construction (Halbach, 2011, p. 211). Moreover, for large stages α before the minimal fixed point is reached, the system *KF* proves sentences that first become true in the model that is constructed at stage α.

To some extent, versions of the revision theory of truth have also been connected with laws of truth (Halbach, 1994; Gupta & Belnap, 1993; Horsten, 2011). In the second part of this article, this research will be carried further. But our attempts will only succeed to a limited extent. We will see that we are not able axiomatically to capture the spirit of the revision theory to the extent that Kripke's theory has been captured.

The axiomatic system *FS* of Friedman & Sheard (1987) is nearly stably true. But we will argue that, from a truth theoretic point of view, *FS* is ultimately not very attractive. The problems for *FS* relate to the phenomenon (discussed in Halbach & Horsten, 2005) that reflection principles cannot be added to *FS* in a natural way, which in turn derives from the fact that *FS* is ω-inconsistent.

Instead, we will concentrate on a system, which we will call *PosFS* ("Positive *FS*"), and which is stably true. We will argue that *PosFS* is a more natural truth theory. *PosFS* is compositional to a high degree, its inner logic coincides with its outer logic, and reflection principles can be added to it in a completely straightforward way.

§2. Two Revision Theoretic notions of truth. The general idea of the Revision Theory of Truth is the following. We start with a classical model for $\mathcal{L}_T$. This model is transformed into a new model again and again, thus yielding a long sequence of classical models for $\mathcal{L}_T$, which are indexed by ordinal numbers. The official notion of truth for a formula of $\mathcal{L}_T$ is then distilled from this long sequence of models.

We only consider models that are based on the standard natural number structure. So all models that we will consider will be of the form $\langle \mathbb{N}, S \rangle$, where $\mathbb{N}$ specifies the domain of discourse and the interpretation of the arithmetical vocabulary, and S specifies the extension of the truth predicate.

For simplicity, let us start with the model

$$\mathfrak{M}_0 = \langle \mathbb{N}, \emptyset \rangle :$$

the model that regards no sentence whatsoever as true (corresponding to a proposal made by Herzberger). Suppose we have a model $\mathfrak{M}_\alpha$. Then the next model in the sequence is defined as follows:

$$\mathfrak{M}_{\alpha+1} = \langle \mathbb{N}, \{\phi \in \mathcal{L}_T \mid \mathfrak{M}_\alpha \models \phi\} \rangle.$$

In other words, the next model is always obtained by putting those sentences in the extension of the truth predicate that are made true by the last model that has already been obtained.

Now suppose that λ is a limit ordinal, and that all models $\mathfrak{M}_\beta$ for $\beta < \lambda$ have already been defined. Then

$$\mathfrak{M}_\lambda = \langle \mathbb{N}, \{\phi \in \mathcal{L}_T \mid \exists\beta\forall\gamma : (\gamma \geq \beta \wedge \gamma < \lambda) \Rightarrow \mathfrak{M}_\gamma \models \phi\} \rangle.$$

In words: we put a sentence ϕ in the extension of the truth predicate of $\mathfrak{M}_\lambda$ if there is a "stage" β before λ such that from $\mathfrak{M}_\beta$ onwards, ϕ is *always* in the extension of the truth predicate. The sentences in the extension of the truth predicate of $\mathfrak{M}_\lambda$ are those that have "stabilized" to the value *True* at some stage before λ.[2]

This yields a chain of models—and a corresponding chain of extensions H_α of the truth predicate—that is as long as the chain of the ordinal numbers. Elementary cardinality considerations (Cantor's theorem) tell us that there must be ordinals α and β such that $\mathfrak{M}_\alpha$ and $\mathfrak{M}_\beta$ are identical. In other words, the chain of models must be periodic. The first α such that $\mathfrak{M}_\alpha$ is recurring in this way in the sequence is called ζ, and the first stage where it re-occurs is called Σ.

On the basis of this long sequence of models, one can then define the notion of *stable truth* for the language $\mathcal{L}_T$. A sentence $\phi \in \mathcal{L}_T$ is said to be stably true if at some ordinal stage α, ϕ enters in the extension of the truth predicate of $\mathfrak{M}_\alpha$ and stays in the extension of the truth predicate in all later models. A sentence $\phi \in \mathcal{L}_T$ is said to be stably false if at some ordinal stage α, ϕ is outside the extension of the truth predicate of $\mathfrak{M}_\alpha$ and stays out forever thereafter. A sentence that is neither stably true nor stably false is said to be *paradoxical*.

Revision theorists have tentatively proposed to identify truth simpliciter with stable truth and falsehood simpliciter with stable falsehood, whilst sentences that never stabilize, such as the liar, are classified as paradoxical. But they hesitate to endorse this identification. Another strong contender for identification with truth simpliciter (falsehood simpliciter) is the slightly more complicated notion of *nearly stable truth* (nearly stable falsehood). A sentence $\phi \in \mathcal{L}_T$ is said to be nearly stably true if for every stage α after some stage β, there is a natural number n such that for all natural numbers $m \geq n$, ϕ is in the extension of the truth predicate of $\mathfrak{M}_{\alpha+m}$. And a sentence $\phi \in \mathcal{L}_T$ is said to be nearly stably false if for every stage α after some stage β, there is a natural number n such that for all natural numbers $m \geq n$, ϕ is outside the extension of the truth predicate of $\mathfrak{M}_{\alpha+m}$. In other words, for this notion of truth we do not care what happens before any fixed finite number of steps after any limit ordinal.

The notions of stable truth and nearly stable truth do not coincide. Consider, for instance, the sentence

$$\forall \phi \in \mathcal{L}_T : \neg T(\phi) \leftrightarrow T(\neg\phi).$$

This sentence is false only at limit stage models in the chain of revision models. Therefore it is nearly stably true, but not stably true.

§3. Determinateness.

3.1. Field's determinacy predicates. It was observed in passing in Welch (2008) that it is possible to effect Field's notion of *determinateness* for his theory of truth (*cf.* Field, 2003, 2008) for the set of truths in a Herzberger revision sequence. We expand on this observation here.

We first summarize Field's notions. Field seeks to express the *defectiveness* of a simple liar sentence Q_0 by the use of a *determinateness operator*. He defines $D(A) \equiv A \wedge \neg(A \rightarrow \neg A)$. In Field (2003, 2008) the construction permits this to be $A \rightarrow (\top \rightarrow A)$. From the evaluations one gives to the $\rightarrow$ operator one can see that we can think of this as "A is true now and was so at the previous stage." This D operation is iterated, and moreover transfinitely. Field in his papers has some difficult discussion on the lengths of these possible hierarchies as iterated along "independent paths." It was shown in Welch (201?)

---

[2] So we disregard, in this article, all other limit rules for revision sequences that have been thought of in the literature.

how to make sense of this in terms of the internally defined prewellorderings of where his sentences' truth values stabilize before his "first acceptable point." (See the discussion also in Welch, 2011.) The latter he calls Δ_0 but we adopt the notation of ζ for this point. It was shown in Burgess (1986) that for a Herzberger sequence starting from the empty hypothesis, that the sequence would repeat at the least ζ for which there was a larger ξ with $L_\zeta \prec_{\Sigma_2} L_\xi$. Indeed this ordinal pair occurs in Field's theory as the first two acceptable points (Welch, 2008). We now enunciate some of Field's desiderata for his determinateness operator, really so as the reader may check that our operator will meet them, and how very close the two behaviors are. An understanding of this discussion is not necessary for our definitions, so the reader may with impunity skip ahead to Section 3.2 if they wish.

Field argues for the desirability of a notion of *determinate truth* as a way of expressing within the language the feeling that somehow the liar is "defective" and that we should have a way of expressing this. Then for him, we see that the simple liar Q_0 has ultimate value $\|Q_0\| = \frac{1}{2}$. (We use Q's to avoid confusion with the levels of the L_α hierarchy, which will become relevant soon.) DQ_0 however has ultimate value 0 and so the determinate truth value of this liar is certainly 0. In terms of D we can form by diagonalization a second liar Q_1 which is stronger. For this liar $\|DQ_1\| = \frac{1}{2}$, but we have $\|DDQ_1\| = 0$. This is generalizable: determinateness operators D^n are created hand-in-hand with strengthened liars Q_n, and on into the recursive transfinite forming Q_α and D^α sequences, and the natural question is how far this can go.

Field's desiderata for such determinateness hierarchies (Field, 2008, p. 256) are (using single bars for semantic values) summarized as:

- $|A| = 1 \Rightarrow |DA| = 1$
- $|A| = 0 \Rightarrow |DA| = 0$
- $0 \prec_d |A| \prec_d 1 \Rightarrow |DA| \prec_d |A|$
- $|A| \preceq_d |B| \Rightarrow |DA| \preceq_d |DB|$

In Field's principle construction of a model, the above $\prec_d$ is the order on the de Morgan algebra of semantic values given by the functional valuations in the first "period" $[\Delta_0, \Delta_1)$ between the first two acceptable points (see his Chapter 17.1). For the most part he considers, and we will too, the three valued ordering of $\{0, \frac{1}{2}, 1\}$. We also have the same periodicity phenomenon (indeed the same period!) in the Herzberger revision sequence as he has for his standard construction over the usual model of arithmetic. We shall just simply think of semantic value $\frac{1}{2}$ as being the "unstable truth value" and write $|A| = \uparrow$ for such, but again for the purposes of comparison with Field's ordering, still think of $\uparrow$ as of intermediate value between 0 and 1.

We have the liar hierarchy:

$$\begin{aligned} Q_0 &= \neg TQ_0 \\ Q_1 &= \neg DTQ_1 \\ Q_2 &= \neg DDTQ_2 \text{ etc.} \end{aligned}$$

And by various possible devices, into the transfinite:

$$Q_\sigma = \neg D^\sigma TQ_\sigma.$$

Then he has:

- For any σ: $\|D^{\sigma+1} Q_\sigma\| = 0 \neq \|D^\sigma Q_\sigma\|$.

Further (top of p. 255) if we analyze the values of Q_σ at stages, then $|Q_0|_\alpha = \frac{1}{2}$ for every α and more generally, Q_σ will have cycles consisting of a $\frac{1}{2}$ followed by a σ-sequence of 1's. However for example for finite k, $D^k Q_\sigma$ has cycles of the same length but first as $\frac{1}{2}$ then followed by k 0's, then 1's.

3.2. Determinacy predicates for the revision theory. It is the aim of this section to demonstrate that these determinateness notions can be decoupled from Field's conditional $\rightarrow$, and defined within the Herzbergerian style theory. We shall see that we easily get similar phenomena if we define for a Herzberger sequence setting a determinateness operator D_h:

$$D_h A = A \wedge TA; \quad D_h^{\sigma+1} A = D_h(D_h^\sigma A); \quad D_h^\lambda A \equiv \forall\sigma < \lambda(TD^\sigma A) \quad \text{for } Lim(\lambda)$$

(using again some as yet unspecified means of formally coding the infinitary conjunction in the limit case of the right hand side above; we shall defer doing this properly until we see how to do it for all possible $\alpha < \zeta$).

The Liar hierarchy Q_σ can be defined as above. The readers can calculate for themselves the behavior of these liars in a typical revision sequence starting out with all sentences having value 0. For example DQ_0 is 0 at every stage. Again for $k < \omega$ the periodicity of Q_k is $k + 1$—it just simply flips back and forth in value every $k + 1$ steps.

One might also point out that after limit stages various levels of the D_h^k are equivalent: let $Lim(\lambda)$, then we have $D_h^k A \in H_{\lambda+j} \Leftrightarrow D_h^m A \in H_{\lambda+j}$ for any $j < k \leq m \leq \omega$. (Again this is not special to Herzberger, but occurs in the Fieldian principal construction too.) The reader may also verify that the desiderata listed for D above hold also for D_h (either in the simplified three valued form described above, or in the de Morgan algebra form described in Field, 2008, chapter 17.1).

One may then ask how long such determinateness hierarchies can be sensibly extended, and we can answer this in entirely analogous manner to that for the Fieldian hierarchy, as was done in Welch (201?). Although the successor stages of the two processes, Herzbergerian and Fieldian are completely different, the common liminf process at limit stages (being some form of strong infinitary rule), means that the analyses are, up to a change in notation, identical in spirit. As the reader may perhaps be loathe to go through the details, we perform this task in Section 3.2.1. However those seeking the full provenance of the arguments and ideas should consult Welch (201?).

We shall see the length for such hierarchies, as for Field, is ζ. The key to doing this are the following uniform definability results:

LEMMA 3.1. (Welch, 201?, Wellordering Lemma) *There is a single uniform recursively enumerable method of defining a wellordering w_β of order type β from H_β for any limit $\beta < \Sigma$. This method is uniform in the sense that it is independent of β.*

This is amplified as follows:

LEMMA 3.2. (Welch, 201?, "Uniform Definability") *There is a single uniform method of arithmetically defining (a set of integers coding) the whole sequence $\langle H_\gamma \mid \gamma < \beta\rangle$ from H_β for any $\beta < \Sigma$. Again this method is uniform in the sense that it is independent of β.*

In the case of a successor $\beta = \gamma + 1 < \Sigma$ we may even assert that there is a single *recursive* function (thus independent of β) $F : \mathbb{N}^2 \longrightarrow \mathbb{N}$, so that if we set

$$\mathcal{H} = \left\{\langle\ulcorner A \urcorner, u\rangle \in \mathbb{N}^2 \mid F(\langle\ulcorner A \urcorner, u\rangle) \in H_\beta\right\}$$

then with w_β the well ordering of type β from the Wellordering Lemma above, and $u \in w_\beta$, then, if u has rank γ in w_β then $\mathcal{H}_u =_{df} \{\ulcorner A \urcorner \mid \langle \ulcorner A \urcorner, u\rangle \in \mathcal{H}\}$ is nothing other than H_γ itself. Thus for such β we have a way not only of defining simply a wellorder of type β from H_β, but we may *recursively* recover the whole prior sequence $\langle H_\gamma \mid \gamma < \beta\rangle$ from knowledge of H_β. Again the method is independent of β. Hence we may think of H_β as always encoding the whole revision sequence up to β.

The idea is that since we can recover the previous sequence from the truth set H_β, we in fact at stage β have a knowledge about when particular sentences stabilize before stage β. This allows us (with the uniformity above) to build a formula $P_\prec(v_0, v_1)$ which when evaluated at stage β, will be true of $\ulcorner A \urcorner, \ulcorner B \urcorner$ if A has stabilized below β before B has. Of course this evaluation, $|P_\prec(\ulcorner A \urcorner, \ulcorner B \urcorner)|_\beta$, then may change later, but the eventual values (in Field's notation $\|P_\prec(\ulcorner A \urcorner, \ulcorner B \urcorner)\|$) will tell us whether we really do have this stabilization or not. We may paraphrase the above as saying that we may work "as if" there were predicate letters $\dot{H}_\alpha$ in the language at stage β for any $\alpha < \beta$: we can refer to the extensions of these predicates with ease. In effect we are using the sentences that stabilize as notations for the ordinal which is the rank of their order of stabilization.

More formally: for a sentence A we may define $\rho(A)$ to be the least ordinal ρ (if it exists) in a revision sequence so that the semantic value of A is constant from stage ρ onwards. We let "$\rho(A)\downarrow$" abbreviate the assertion that $\rho(A)$ is defined.

We may define in the language L_T a *prewellordering* $\prec$ of sentences of stabilizing truth value: we set $P_\prec(\ulcorner A \urcorner, \ulcorner B \urcorner)$ if and only if $\rho(A) < \rho(B)$, where $\ulcorner A \urcorner$ is an integer Gödel code for A. (It has to be shown that we can do this and that $P_\prec$ is given by an L_T formula.) The ordering $\preceq$ derived from $\prec$ is a *pre*wellordering since many sentences A may stabilize at the same ordinal. We shall continue to use the notation of $\|A\|$ but now for the stable semantic value of the sentence A (if it exists). Thus $\|A\| = 1$ (or 0) $\leftrightarrow$ $\ulcorner A \urcorner$ (respectively $\ulcorner \neg A \urcorner$) $\in H_\zeta$.

LEMMA 3.3. *There are formulae $P_\preceq(v_0, v_1)$, $P_\prec(v_0, v_1)$ in L_T so that for any sentences $A, B \in L_T$, we have*

$$\begin{aligned}
\|P_\prec(\ulcorner A \urcorner, \ulcorner B \urcorner)\| &= 1 \text{ iff } \rho(A)\downarrow, \rho(B)\downarrow \text{ and } \rho(A) < \rho(B);\\
&= 0 \text{ iff } \rho(A)\downarrow, \rho(B)\downarrow \text{ and } \rho(A) \geq \rho(B);\\
&= \uparrow \text{ otherwise.}
\end{aligned}$$

(And similarly for the formula $P_\preceq$.)

We abbreviate $A \prec B$ for $\|P_\prec(\ulcorner A \urcorner, \ulcorner B \urcorner)\| = 1$ *etc.* Then, if $\|A\| = 1$ (or 0) say, then $\{B : B \prec A\} = \{B : \|P_\prec(\ulcorner A \urcorner, \ulcorner B \urcorner)\| = 1\}$ is a prewellordering of order type some ordinal $\xi < \zeta$. It is less than ζ since, recall, a sentence's eventual status as stably false/true or unstable is decided by, and is reflected precisely at, ordinal stage ζ. The ordinal is highly closed (very "admissible") and this ensures the length of the prewellorder is no greater than ζ (by analogy with the recursive ordinals as all having length less than the height of the least admissible set containing $\omega + 1$, namely ω_1^{ck}). We let Field($\prec$) denote the set of sentences stabilizing on 0 or 1. The next lemma shows how long these prewellorderings can be:

LEMMA 3.4. *For any $\xi < \zeta$ there is a sentence $A = A_\xi$ in Field($\prec$) with the order type of $\{B \mid B \prec A\}$ equalling ξ.*

However this is the extent of the internally definable hierarchies:

LEMMA 3.5. *Let* $Q(v_0, v_1)$ *be a formula of* L_T. *Define* $n \prec_Q m$ *if* $\|Q(n,m)\| = 1$. *Suppose* $\prec_Q$ *is a prewellordering, and further that for any* $m \in Field(\prec_Q)$, *for any* $n \in \mathbb{N}$ $Q(n,m)$ *has a stabilized value. Then* $ot(\prec_Q) \leq \zeta$.

We may now define internal hierarchies of iterated determinateness along initial segments of $\prec$ given by the sets $\{B : B \prec A\}$. We may define for *any* sentence C:

$$D^C_h A \equiv \forall B \prec C \forall y (y = \ulcorner D^B_h A \urcorner \rightarrow Ty).$$

For $C \in Field(\preceq)$ this defines a "genuine" internal determinateness hierarchy of length $\rho(C)$.

The definition makes sense for a general C whether or not it is in $Field(\preceq)$. However if $C \in Field(\preceq)$ we may show:

LEMMA 3.6. *If* $C \in Field(\preceq)$ *then for all* B *either "*$B \preceq C$*" or "*$\neg B \preceq C$*" is in* H_ζ.

3.2.1. Proofs of the Lemmata. *First Lemma 3.3.* This is similar to the proof offered in Welch (201?). By the Uniform Definability Lemma there is a single arithmetical formula Φ that defines over any $\langle \mathbb{N}, H_\beta \rangle$ ($\beta < \Sigma$) a wellorder of type β together with the associated previous H-sets $\langle H_\alpha \mid \alpha < \beta \rangle$.

Thus whether a particular sentence A is stably 0, is then translatable into a two valued arithmetic statement in the language of arithmetic augmented by a symbol for H_β, that is, or is not, true in $\langle \mathbb{N}, H_\beta \rangle$. Let $|\ulcorner A \urcorner|_\alpha$ denote the $0/1$ value that sentence A has at stage α, that is, as to whether $\ulcorner A \urcorner \in H_\alpha$ or not. Let $X(x)$ be the set-theoretic statement: "$\forall \alpha \exists \beta > \alpha |x|_\beta \neq |x|_\alpha$" which expresses that x is the Gödel number of a sentence which has an unstable semantic value. Now translate this, using our Uniform Definability Lemma, as a one place arithmetic predicate $A_X(v_0)$. We assume this is effected in such a way so that $\{\ulcorner B \urcorner \mid \langle \mathbb{N}, H_\beta \rangle \models A_X(\ulcorner B \urcorner)\}$ is the set of sentences unstable below β.

Note that $\|A_X(x)\| = 0 \leftrightarrow \rho(x) \downarrow$. If $\beta = \delta + 1$ then trivially $\langle \mathbb{N}, H_\beta \rangle \models \neg A_X(n)$ for any sentence with code n. However if $Lim(\beta)$ then $\langle \mathbb{N}, H_\beta \rangle \models A_X(n)$ will occur if n is unstable below β. In that case

$$|A_X(n)|_\beta = 1 \wedge |T\ulcorner A_X(n) \urcorner|_{\beta+1} = 1.$$

In conclusion:

$$\rho(x) \downarrow \leftrightarrow \|TA_X(x)\| = 0 \leftrightarrow \|A_X(x)\| = 0$$

Just as in Welch (201?), let $\Psi_\preceq(x, y)$ be:

$X(x) \vee [\neg X(x) \wedge \neg X(y) \wedge$ *if* α_x, α_y *are least so that*

$$\forall \beta \geq \alpha_x \forall \gamma \geq \alpha_y \left(|x|_\beta = |x|_{\alpha_x} \wedge |y|_\gamma = |y|_{\alpha_y} \right) \textit{ then } \alpha_x \leq \alpha_y].$$

Let $A_{\Psi_\preceq}(v_0, v_1)$ be the translation of $\Psi_\preceq(x, y)$ and let $P_\preceq(x, y) \equiv A_{\Psi_\preceq}(x, y)$ be the corresponding L_T formula. We check that $P_\preceq$ is as demanded by the lemma.

Claim:

$$\begin{aligned} \|P_\preceq(\ulcorner A \urcorner, \ulcorner B \urcorner)\| &= 1 \text{ iff } \rho(A) \downarrow, \rho(B) \downarrow \wedge \rho(A) \leq \rho(B) \\ &= 0 \text{ iff } \rho(A) \downarrow, \rho(B) \downarrow \wedge \rho(A) > \rho(B) \\ &= \uparrow \text{ otherwise.} \end{aligned}$$

Proof of Claim: Note that the first line is straightforward:

$$\|P_\preceq(x, y)\| = 1 \leftrightarrow \|A_{\Psi_\preceq}(x, y)\| = 1 \leftrightarrow \|A_X(x)\| = \|A_X(y)\| = 0 \wedge \rho(x) \leq \rho(y).$$

For the second line suppose first $\|P_{\preceq}(x,y)\| = 0$. Then x is stable since otherwise $\|x\|$, $\|A_X(x)\| =\uparrow$, and this would imply $\|P_{\preceq}(x,y)\| =\uparrow$. Then for arbitrarily large $\gamma \in (\rho(x), \zeta)$ we have that, if $\tilde{A}_\alpha(y)$ is the translate of "α_y exists" then $\langle \mathbb{N}, H_\gamma \rangle \models \tilde{A}_\alpha(y)$. (Consider, e.g., any successor $\gamma = \delta + 1$, then $\alpha(y)$ is defined below γ and is $\leq \delta$—it may only be δ itself if y changed semantic value unboundedly in δ with $Lim(\delta)$.) If $\langle \mathbb{N}, H_\gamma \rangle \models \tilde{A}_\alpha(y)$ and also α_y as defined over $\langle \mathbb{N}, H_\gamma \rangle$ were greater than or equal to $\rho(x)$ we should have $\langle \mathbb{N}, H_\gamma \rangle \models A_{\Psi_{\preceq}}(x,y)$. However $\|A_{\Psi_{\preceq}}(x,y)\|$ is supposed to be 0, that is, to have a zero value on a final segment below ζ. So for such γ we always must have $\alpha_y < \rho(x)$. However that implies $\rho(y)\downarrow \wedge\, \rho(y) < \rho(x)$.

The converse is straightforward. Hence $\|P_{\preceq}(x,y)\| =\uparrow$ in the remaining cases. The definition of $P_{\prec}(x,y)$ is done analogously. QED Lemma 3.3.

Proof of Lemma 3.4. It suffices to show that $\zeta_0 =_{df} ot(\prec) = \zeta$. Note first that $\zeta_0 \leq \zeta$ since by definition of ζ it is the least point where the revision sequence starts to cycle, that is, any sentence that is going to stabilize will do so by stage ζ. We show that $\zeta_0 \geq \zeta$. We summarize the idea as follows: since we have a recursive function $G : \mathbb{N} \to \mathbb{N}$ with the Σ_2-Theory of L_ζ the preimage under G of H_ζ, part of that theory contains the sentences "$n \in Field(w_\zeta)$" and "$n <_{w_\zeta} m$" where w_ζ is the uniformly $\Sigma_2^{L_\zeta}$ well order of type ζ. Since such set-theoretic sentences "settle down" in order type ζ the corresponding arithmetical sentences $G(\ulcorner n \in Field(w_\zeta)\urcorner)$ settle down into H_ζ also in the same order type. This is worked out in detail below.

As intimated, we have a canonical $\Sigma_2^{L_\zeta}$ definable partial function $g_\zeta; \omega \longrightarrow \zeta$ which is onto, for any α if n_α is such that $g_\zeta(n_\alpha) = \alpha$, the statement Φ_α: "$n_\alpha \in dom(g)$" is part of the Σ_2-theory of L_ζ, which itself is true in some $L_{\rho(\alpha)}$ onwards. We shall show that there is a ζ-long sequence, S, of α so that for $\alpha < \alpha' \in S$, $\rho(\alpha) < \rho(\alpha')$. Assuming for the moment this is shown, T^2_ζ, the Σ_2-theory of L_ζ is recursive in H_ζ, (L_ζ being a model of Σ_1-Separation); let G be (1-1) and recursive witnessing that $T^2_\zeta \leq_1 H_\zeta$. There is then some $A_\alpha \in H_\zeta$ so that $G(\Phi_\alpha) = A_\alpha$. The value of $|A_\alpha|_\beta$ is then stable from $\rho(\alpha)$ onwards. As the $\rho(\alpha)$'s form a ζ-sequence unbounded in ζ, this will establish that $\zeta_0 \geq \zeta$ as required.

We take $S = S^1_\zeta =_{df} \{\alpha \mid L_\alpha \prec_{\Sigma_1} L_\zeta\}$. By the reflection property that defines ζ as the least such that there is $\Sigma > \zeta$ with $L_\zeta \prec_{\Sigma_2} L_\Sigma$, one may show that S is unbounded in ζ and has order type ζ. We use the definition of $\rho(\alpha)$ from Φ_α, in the last paragraph.

Claim For $\alpha \in S$, $\alpha' > \rho(\alpha) \geq \alpha$ where α' is the least element of S above α. Hence for $\alpha < \alpha' \in S$, $\rho(\alpha) < \rho(\alpha')$.

Proof of Claim: Let α, α' be as asserted in the Claim. We reemphasize:

(1) The definition of g_ζ is uniform in ζ, meaning that g_β (for limit β) is defined over L_β by the same definition. There is thus some $\Psi(v_0, v_1) \equiv \exists u \forall v \chi(u, v, v_0, v_1)$ with χ Σ_0, so that $L_\beta \models \Psi(n, \beta)$ iff $g_\beta(n) = \beta$, with the same $\Psi(v_0, v_1)$ defining a partial function g_β over each such β for $Lim \cap \Sigma$.

Moreover:

(2) $L_\alpha \models$"$g_\alpha(n) = \beta$" $\Longrightarrow L_\zeta \models$"$g_\zeta(n) = \beta$."

Proof of (2): This is because $\alpha \in S^1_\zeta$, and so the Σ_2 formula $\Psi(n, \beta) \equiv g_\alpha(n) = \beta$ persists up to L_ζ. Q.E.D.(2)

This directly implies:

(3) $g_\zeta \cap (\omega \times \alpha) = g_\alpha$ (for $\alpha \in S$).

If $g_\zeta(n_\alpha) = \alpha$, then this statement cannot have become true before α (since $L_\alpha \models$ "$g_\alpha(n_\alpha)$ is undefined"). Hence $\rho(\alpha) \geq \alpha$. However by (3)

$$L_{\alpha'} \models \text{“}g_{\alpha'}(n_\alpha) = \alpha\text{”}$$

and by (2) this is stabilized. Hence $\alpha' > \rho(\alpha)$. Q.E.D.(Claim) & Lemma 3.4

Proof of Lemma 3.6: (This proof is almost verbatim that of Welch, 201?, Lemma 16 but is again included for completeness.) Note that $B \preceq C_0$ implies

$$\|\text{“}\neg\exists\sigma\exists\rho[\sigma > \rho = \rho(C_0) \wedge |B|_\sigma \neq |B|_{\sigma+1}]\text{”}\| = 1,$$

whilst $B \not\preceq C_0$ implies that this stable value is 0. Using our translations outlined above, the statement within quotes in the last displayed line, has a translation into arithmetic about the $\langle \mathbb{N}, H_\beta \rangle$. Thus, "$\rho = \rho(C_0)$" can be written out using the "stability" formula $X(v_0)$ and corresponding $A_X(v_0)$. Then questions concerning whether $B \preceq C_0$ or not can be answered by consulting H_ζ. Q.E.D.

It is tedious to check, but not very hard to verify, that the same results hold for the revision-theoretic notion of nearly stable truth. This was to be expected, given the fact that the complexity of the notion of nearly stable truth is the same as the complexity of the notion of stable truth.

3.3. Beating the determinateness hierarchy. The foregoing may suggest that there exists a situation of "Mutual Assured Destruction" between strengthened liar sentences on the one hand, and the hierarchy of indeterminateness predicates on the other hand. But this is not quite correct. There are super-liar sentences such that their paradoxicality is not stably attested by any of the indeterminacy predicates in the revision-theoretic hierarchy. Intuitively, what happens is this. Liar-like sentences change their truth value periodically. Liar-like sentences that have a large period can only be "caught" by indeterminacy predicates higher up in the hierarchy. But there are liar-like sentences of which the period is so long that they escape the revision-theoretic indeterminacy hierarchy altogether.

In the Fieldian setting, the situation is similar. Even though every time a super-liar "diagonalizes out" of a given indeterminacy predicate, it is captured by an indeterminacy predicate of the next higher order. Nonetheless, an ineffable liar exists that escapes all predicates in the Fieldian indeterminacy hierarchy. This reinforces our conclusion that as far as the treatment of super-liar sentences is concerned, Field's theory does not hold any advantages over the revision theory of truth.

Now, the mathematical models that Field produces should not be identified with his theory of truth, since ultimately his theory is meant to be presented in axiomatic terms. Nonetheless, his models are intended to serve as models in which we see the logical behavior of truth and determinacy in action. One of the two selling points of Field's truth theory is that it claims to solve the problem of the strengthened liar paradox. (The other is that it specifies a way in which the unrestricted Tarski-biconditionals can be taken to hold.) The phenomenon of ineffable liars therefore does show that Field has more work to do before we can be convinced that the problem of the strengthened liar has been laid to rest.[3]

Here we analyze the situation in the simpler but relevantly similar setting of the revision theory. There exist also in the revision theory no "periodic" sentences with periods less than

---

[3] For more on the analogue in the Fieldian setting, see Welch (201?).

Σ that escape the indeterminacy hierarchy. This is because any "course-of-values-periodic" sentence with a period $< \Sigma$ actually has one with period $< \zeta$ (by the reflecting properties of L_ζ).) But there is a sentence σ so that for any $\delta < \Sigma$ there are $\gamma > \rho > \delta$ with σ having the same value in the interval $[\gamma, \gamma + \rho)$ and the opposite value in $[\gamma + \rho, \gamma + \rho \times 2)$. Such a sentence cannot be dominated (= made "determinately 0") by any determinateness predicate in the internal hierarchy: it is too "sporadic". (Although it does have a period: namely Σ itself.) Hence these sporadic sentences form a subclass of the unstables not in H_ζ which outwit all the determinateness predicates as defined above. In a precise sense it is these sporadic sentences that "diagonalize out" of the sets internally definable by using $\langle \mathbb{N}, H_\zeta \rangle$: for $C \in Field(\preceq)$ those A with $\|D_h^C(A)\| = 0$ form an internally definable class (meaning that there is a formula $\varphi(v_0)$ which has stable value 1 for $\varphi(A)$ iff "$D_h^C(A) = 0$" has stable value 1—this is only repeating the text). So even though such A are unstable, we can internally categorize them so to speak. The sporadics ineffably defy such definable defectiveness categorizations.

Let us now look at the details.

PROPOSITION 3.7. *There are sentences $C \in \mathcal{L}_T$ so that for any determinateness predicate D^B with $B \in Field(\preceq)$ $\|D^B(Q_C)\| =\uparrow$, that is $D^B(Q_C)$ is unstable. Thus the defectiveness of Q_C is not measured by any such determinateness predicate definable within the $\mathcal{L}_T$ language.*

Proof: Further, as $\mathbb{N} \in L_{\omega+1}$ and the successive levels of the revision construction are performed using very absolute processes, we may consider running the construction "inside of" the L-hierarchy. The ordinals ζ, Σ are highly closed, and in fact highly admissible. We set $ADM^+ = ADM \cap ADM^*$ to be the class of admissible limits of admissible ordinals, We may define predicates in the language of set theory that give us the range of semantic values of sentences along the Herzberger iteration. So that, if $\tau \in ADM^+$ then $(|A|_\gamma = i)_{L_\tau} \leftrightarrow |A|_\gamma = i$, (in other words that $A \in H_\gamma \leftrightarrow (A \in H_\gamma)^{L_\tau}$. Thus the construction is absolute to L_τ. We can see readily what happens for small ordinal iterations of D: if $\alpha < \sigma$ then $D^\alpha(Q_\sigma)$ cycles through an α-sequence of 0's, and then a tail of 1's making a σ-sequence altogether, before looping around again. $D^\sigma(Q_\sigma)$ will cycle through a σ-sequence of 0's before repeating; finally $D^{\sigma+1}(Q_\sigma)$ will be always 0. Hence $\|D^{\sigma+1}(Q_\sigma)\| = 0$, and thus the "defectiveness" of Q_σ is affirmed by this sentence. Essentially the same picture is intended for these extended operators, where now α, σ *etc.* are replaced by sentences $B, C, \ldots$ as notations.

(1) *There are ordinals $\Sigma > \gamma > \xi > \zeta$ and a sentence C with $\gamma \in ADM^+$ and $L_\gamma \models$ "$\rho(C) = \xi$."*

Proof: If not, then the following is true in L_Σ:

$$y = \zeta \leftrightarrow y \in ADM^+ \wedge L_y \models \text{“}\forall \xi \exists C(\rho(C) = \xi)\text{”} \wedge$$
$$\wedge \forall y' \in ADM^+(y' > y \longrightarrow L_{y'} \models \text{“}\forall C(\rho(C) \downarrow \longrightarrow \rho(C) \leq y).\text{”}$$

Being in ADM^+ is a Δ_1 notion, as are the satisfaction relations involving $L_y, L_{y'}$. We note that $\zeta \in ADM^+$. The second conjunct holds since $rk(\preceq) = \zeta$, and all $B \in Field(\preceq)$ have stabilized by stage ζ. The last conjunct is our hypothesis. However this would imply that ζ is Π_1 definable (by the above definition) without using any other parameters in L_Σ). But it is not: only sets in L_ζ can be Σ_2 definable without parameters in L_Σ (since $L_\zeta \prec_{\Sigma_2} L_\Sigma$). It particular ζ itself is not so definable. Q.E.D.(1)

Let C be as guaranteed in (1). Let $\bar{\zeta} < \zeta$ be arbitrary. Then we have (as a restatement, and weakening, of the above):

(2) $L_\Sigma \models$"$\exists\gamma \in ADM^+(L_\gamma \models \rho(C) > \bar{\zeta})$."

By Σ_1-elementarity then:

(3) $L_\zeta \models$"$\exists\gamma \in ADM^+(L_\gamma \models \rho(C) > \bar{\zeta})$."

But $\bar{\zeta}$ was arbitrarily large below ζ, thus, in fact:

(4) $L_\zeta \models$"$\forall\bar{\zeta}\exists\gamma > \bar{\zeta}(\gamma \in ADM^+ \wedge L_\gamma \models \rho(C) > \bar{\zeta})$."

The claim is that, staying with this C, that it satisfies the proposition. Pick any $B \in Field(\preceq)$. It suffices to show that

(5) $\forall\bar{\tau} < \zeta\exists\tau > \bar{\tau}(\tau < \zeta \wedge |D^B(Q_C)|_\tau \neq 0)$.

Proof (5): Taking $\bar{\tau}$ any ordinal greater than $\rho(B)$, then by (3) (with $\bar{\tau}$ as $\bar{\zeta}$ there) there is $\gamma \in ADM^+$ with $L_\gamma \models \rho(C) > \rho(B)$. By choice, γ is an admissible limit of admissibles, so γ iterations of the Fieldian construction can be effected inside L_γ. But then inside L_γ we see the usual picture of the cycling semantic values of $0, 0, \ldots$ (for $\rho(B)$ steps) and 1's for $\rho(C) - \rho(B)$ steps, then repeating this pattern. Consequently, with $\tau = \gamma$ we see $|D^B(Q_C)|_\tau \neq 0$. Q.E.D.(5) & Proposition.

In fact we can say a little more about such a C: (4) is a Π_2 sentence about C, true in L_ζ and so goes up to be true in L_Σ. So for such a C, it has arbitrarily large $\preceq$-rank, but locally in varying L_γ. One may call such a C *sporadic*. The nonstabilizing sentences in Field's model are of two kinds: those that exhibit a periodic behavior with some fixed period $\xi < \zeta$, (and for every $\xi < \zeta$ there will be such) and the sporadics like C, which have no periodic behavior at all below Σ: if we want to assign a "period" to C it has to be Σ itself.

§4. Principles of nearly stable truth. Now we turn to the sentences which the Revision Theory regards as true. We have seen before that the Revision Theory offers two alternatives. Either truth is to be identified with stable truth, or truth should be identified with nearly stable truth. We first consider the second alternative.

Friedman and Sheard have proposed an axiomatic theory of self-referential truth which is called *FS* (Friedman & Sheard, 1987). Friedman and Sheard gave a slightly different list of axioms (and they did not call their system *FS*), but the following list is equivalent to their system:[4]

- FS1 PA^T, which is Peano Arithmetic with occurrences of T allowed in the induction scheme;
- FS2 $\forall$ atomic $\phi \in \mathcal{L}_{PA} : T(\phi) \leftrightarrow val^+(\phi)$, where val^+ defines atomic arithmetical truth;
- FS3 $\forall\phi \in \mathcal{L}_T : T(\neg\phi) \leftrightarrow \neg T(\phi)$;
- FS4 $\forall\phi, \psi \in \mathcal{L}_T : T(\phi \wedge \psi) \leftrightarrow T(\phi) \wedge T(\psi)$;
- FS5 $\forall\phi(x) \in \mathcal{L}_T : T(\forall x\phi(x)) \leftrightarrow \forall t T(\phi(t/x))$.

[4] This formulation of *FS* is due to Halbach: see Halbach (1994). In the interest of readability, we are somewhat sloppy with notation here. The correct notation is explained in Halbach (2011, Part I, Section 5).

Moreover, *FS* contains two extra rules of inference, which are called *Necessitation* (NEC) and *Co-Necessitation* (CONEC), respectively:

NEC From a proof of ϕ, infer $T(\phi)$;
CONEC From a proof of $T(\phi)$, infer ϕ.

Let us consider the axioms and rules of *FS*. The compositional axioms of *FS* show that it seeks to reflect the intuition of the *compositionality* of truth in a type-free setting. In this sense, *FS* can be seen as a natural extension of the typed compositional theory of truth. In fact, the *axioms* are exactly like the axioms of the typed compositional theory of truth,[5] except that in *FS* the compositional axioms quantify over the entire language of truth instead of only over $\mathcal{L}_{PA}$. But if we disregard the rules of inference NEC and CONEC, this does not help us in any way in proving *iterated* truth statements. The reason is that the truth axiom for atomic sentences only quantifies over atomic *arithmetical* sentences.

FS is the result of maximizing the intuition of the compositionality of truth. Nevertheless, the truth of truth attribution statements is in *FS* only in a weaker sense compositionally determined than the truth of other statements. For *FS* only claims that if a truth attribution has been *proved*, then this truth attribution can be regarded as true (and conversely), whereas for a conjunctive statement, for instance, *FS* makes the stronger hypothetical claim that *if* it is true, then both its conjuncts are true also (and conversely). But it is necessarily that way. If we replace NEC and the CONEC by the corresponding *axiom schemes*, an inconsistent theory results.

Furthermore, we have:

PROPOSITION 4.1. *The theorems of FS are all nearly stably true.*

Proof. This is established by first showing that all the axioms of *FS* are nearly stably true, and by subsequently showing that the nearly stable truths are closed under the inference rules $\phi \Rightarrow T(\phi)$ and $\phi \Rightarrow T(\phi)$. This is routine. □

Proposition 4.1 itself is proved in Gupta & Belnap (1993, p. 222). It entails that *FS* is consistent (for the nearly stable truths are consistent) and indeed arithmetically sound (all the models in the nearly stable truth-sequence are based on the natural numbers). These facts are not new: the former fact is already proved in Friedman & Sheard (1987); a proof of the latter fact is given in Halbach (1994). Halbach (1994) in fact gives an exact computation of the arithmetical strength of *FS*:

THEOREM 4.2. *The arithmetical theorems of FS coincide with the first-order arithmetical consequences of ramified analysis up to stage* ω ($RA_{<\omega}$).

It follows from the main result of McGee (1992) that *FS* is ω-inconsistent. To this end, we consider the sentence γ such that

$$PA \vdash \gamma \leftrightarrow \exists n > 0 : \neg T^n \gamma,$$

which is obtained as an application of the diagonal lemma. We see that:

LEMMA 4.3. $FS \vdash T\gamma \rightarrow \gamma$

Proof. We reason in *FS*. Suppose $T\gamma$, that is, $T\exists n > 0 : \neg T^n \gamma$. By the compositional axioms of *FS*, this entails $\exists n > 0 : \neg T T^n \gamma$, that is, $\exists n > 0 : \neg T^{n+1} \gamma$. And this in turn entails $\exists n > 0 : \neg T^n \gamma$. □

[5] See Horsten (2011, Chapter 6).

Using this lemma, it is easy to see that:

THEOREM 4.4 (McGee). *FS is ω-inconsistent.*

Proof. We reason in *FS*. Suppose $\neg\gamma$. Then $\forall n > 0 : T^n\gamma$. Therefore $T\gamma$, and by the previous lemma, we obtain γ, which is a contradiction. So $FS \vdash \exists n > 0 : \neg T^n\gamma$. But by repeated application of the Necessitation rule, *FS* then also proves $T\gamma, T^2\gamma, T^3\gamma, \ldots$. □

Combining this with Proposition 4.1, this yields (Gupta & Belnap, 1993, pp. 225–227):

COROLLARY 4.5. *The class of nearly stable truths is ω-inconsistent.*

It is often said that when one accepts a theory T, then one is implicitly committed to accepting the soundness of T. In other words, when one accepts T, then one is implicitly committed to the global reflection principle for T.

It goes virtually without saying that the *arithmetical* global reflection principle

$$\forall\phi \in \mathcal{L}_{PA} : Bew_{FS}(\phi) \rightarrow T\phi$$

can be consistently added to *FS* (and it increases the arithmetical strength of *FS*). Indeed, this reflection principle is made true by all models in revision sequences, except perhaps for the initial model (if we, e.g., start with the empty extension for T). And this process of adding an arithmetical reflection principle can be iterated in the familiar way.

But, somewhat remarkably, its truth-theoretic generalization

$$\forall\phi \in \mathcal{L}_{T} : Bew_{FS}(\phi) \rightarrow T\phi,$$

which is called the *global reflection principle* for *FS*, cannot be consistently added to *FS* (Halbach & Horsten, 2005, p. 213):

PROPOSITION 4.6. *FS plus the global reflection principle for FS is inconsistent.*

Proof. We have seen how $FS \vdash \exists n \neg T^n\gamma$. Now *FS* (indeed, already *PA*) proves also that $\forall n\, Bew_{FS}(T^n\gamma)$. So, by the global reflection principle, *FS* concludes that $\forall n T T^n\gamma$. From this, *FS* obtains $\forall n T^n\gamma$, a contradiction. □

This means that *FS* is a theory that is not naturally extendible by means of reflection principles. One can consistently extend it by means of the modified reflection principle

$$\forall\phi \in \mathcal{L}_T \exists n : T^n[Bew_{FS}(\phi) \rightarrow T\phi]$$

As the reader can readily verify, this modified reflection principle is nearly stably true. But it is not a very natural principle. So it is hard to escape the conclusion that the system *FS* is not open-ended in desirable ways. This makes *FS* rather unattractive as a truth theory, despite the defence that Halbach & Horsten (2005) have tried to give of this system.

§5. Principles of stable truth. The compositional axioms fail at limit stages in the sequences that build up nearly stable truth, so *FS* does not belong to the stable truths (Gupta & Belnap, 1993, p. 222). Instead, the stable truths contain a T-positive theory, which we will call *PosFS* (Horsten, 2011, Chapter 8):[6]

PFS1 PA^T;
PFS2 $\forall$ atomic $\phi \in \mathcal{L}_{PA} : T(\phi) \leftrightarrow val^+(\phi)$;

[6] For an early attempt to axiomatize stable truth, see Turner (1990).

PFS3 $\forall$ atomic $\phi \in \mathcal{L}_{PA} : T(\neg\phi) \leftrightarrow val^{-}(\phi)$, where $val^{-}(\phi)$ is an arithmetical formula that defines the atomic arithmetical falsehoods;
PFS4 $\forall\phi, \psi \in \mathcal{L}_T : T(\phi \wedge \psi) \leftrightarrow (T(\phi) \wedge T(\psi))$;
PFS5 $\forall\phi, \psi \in \mathcal{L}_T : (T(\neg\phi) \vee T(\neg\psi)) \rightarrow T(\neg(\phi \wedge \psi))$;
PFS6 $\forall\phi, \psi \in \mathcal{L}_T : (T(\phi) \wedge T\neg(\phi \wedge \neg\psi)) \rightarrow T(\psi)$;
PFS7 $\forall\phi(x) \in \mathcal{L}_T : \exists t T(\neg\phi(t/x)) \rightarrow T(\neg\forall x\phi(x))$;
PFS8 $\forall\phi \in \mathcal{L}_T : \neg(T(\phi) \wedge T(\neg\phi))$ (CONS);
NEC;
CONEC.

It is clear that *PosFS* is a subtheory of *FS*. It is also clear that *PosFS* is not a subtheory of Feferman's theory *KF*: the sentence $T(\lambda \vee \neg\lambda)$, where λ is the liar sentence, is provable in *PosFS* but is not a theorem of *KF*.

An induction on the length of proofs teaches us that:

THEOREM 5.1. *PosFS is stably true.*

Indeed, we have seen in Section 2 that the axiom stating that negation commutes with the truth predicate ($FS3$) is not stably true. This motivates us to "positivize" *FS* in the same way as *KF* can be seen as resulting from a "positivization" of the unrestricted type-free compositional theory of truth (which is of course inconsistent). However, we see that not all of the "positive" *FS*-axioms are stably true. In particular, the converse directions of PFS5 and PFS7, as well as the principle FS5 are not stably true. Therefore they are not included in the list of axioms of *PosFS*.

One also has:

THEOREM 5.2. *PosFS proves no more arithmetical statements than PA.*

Proof. Given a derivation in *PosFS* there is some finite k such that all formulae to which NEC was applied have complexity at most Σ^0_k. As FS5 is not present in the axiomatization of *PosFS* it suffices to successively interpret occurrences of the predicate T in this derivation as Tr_k, the arithmetical truth predicate for Σ^0_k formulae, reducing the derivation to one wholly within *PA*. Since this translation leaves arithmetical formulae unchanged, we deduce that *PosFS* has no more arithmetical theorems than *PA*. □

5.1. Reflection principles. Unlike *FS*, *PosFS* is consistent with its global reflection principle: Since *PosFS* is stably true, so is the statement

$$\forall\phi \in \mathcal{L}_T : Bew_{PosFS}(\phi) \rightarrow T\phi;$$

it becomes true at stage ω. In order to gauge the strength of this principle over *PosFS*, we compare it to arithmetical reflection principles.

Over arithmetical theories there are two natural candidates for reflection principles:

- *Local Reflection* $\mathrm{Rfn}_{\mathcal{L}}(\mathrm{S})$*:* $Bew_{\mathrm{S}}\,\phi \rightarrow \phi$ for each sentence ϕ of $\mathcal{L}$;
- *Uniform Reflection* $\mathrm{RFN}_{\mathcal{L}}(\mathrm{S})$*:* $\forall x : Bew_{\mathrm{S}}\ulcorner\phi(\dot{x})\urcorner \rightarrow \phi(x)$ for each formula $\phi(x)$ of $\mathcal{L}$.

Provided S is a consistent theory, neither principle is derivable in S for regular choices of $\mathcal{L}$ (e.g., $\mathcal{L}$ contains Δ_0). Moreover, the uniform reflection principle is proof-theoretically stronger than local reflection since $\mathrm{S} + \mathrm{RFN}_{\mathcal{L}}(\mathrm{S}) \vdash Consis(\mathrm{S} + \mathrm{Rfn}_{\mathcal{L}}(\mathrm{S}))$ for any theory S. Other forms of reflection include the schema $(\forall x\ Bew_{\mathrm{S}}\ulcorner\phi(\dot{x})\urcorner) \rightarrow \forall x\phi$ and the rule

$$\textit{from a proof of } \forall x : Bew_S\ulcorner\phi(\dot{x})\urcorner, \textit{ infer } \forall x\phi$$

both of which are equivalent to uniform reflection.[7]

Let $\mathrm{GRP}_{\mathcal{L}}(S)$ be the global reflection principle for the theory S over $\mathcal{L}$, that is

$$\forall\phi \in \mathcal{L} : Bew_S\,\phi \rightarrow T\phi. \tag{$\mathrm{GRP}_{\mathcal{L}}(S)$}$$

While the global reflection principle (stated for a theory S) is the natural truth-theoretic version of the local reflection principle, under mild assumptions it is also a generalization of uniform reflection. For example, let $\mathrm{Out}_{\mathcal{L}}$ denote the schema $T\phi(\dot{x}) \rightarrow \phi(x)$ for ϕ from $\mathcal{L}$, and $\mathrm{Out}^-_{\mathcal{L}}$ the restriction to the case where ϕ is a sentence. Then

$$\mathrm{S} + \mathrm{Out}^-_{\mathcal{L}} + \mathrm{GRP}_{\mathcal{L}}(S') \vdash \mathrm{Rfn}_{\mathcal{L}}(S')$$
$$\mathrm{S} + \mathrm{Out}_{\mathcal{L}} + \mathrm{GRP}_{\mathcal{L}}(S') \vdash \mathrm{RFN}_{\mathcal{L}}(S')$$

for any theories S, S'. A more conservative approach would be to replace $\mathrm{Out}_{\mathcal{L}}$ by CONEC; again whether or not parameters are permitted makes an important difference.

Let CONEC^+ denote the rule of co-Necessitation with parameters, that is, the rule

CONEC^+ From a proof of $\forall x : T\phi(\dot{x})$, infer $\forall x\phi$.

PROPOSITION 5.3. *Suppose* S *is any theory formulated in the language* $\mathcal{L}_T \cup \mathcal{L}$. *Then* $\mathrm{S} + \mathrm{CONEC}^+$ *is a subtheory of* S + CONEC + FS5 *and*

$$\mathrm{S} + \mathrm{CONEC}^+ + \mathrm{GRP}_{\mathcal{L}}(S') \vdash \mathrm{RFN}_{\mathcal{L}}(S').$$

Proof. The first part is straightforward as FS5 allows applications of CONEC with parameters to be replaced by applications of parameterless CONEC.

For the second part let S^+ denote the theory $\mathrm{S} + \mathrm{CONEC}^+ + \mathrm{GRP}_{\mathcal{L}}(S')$ and suppose $S^+ \vdash \forall x : Bew_{S'}\,\phi(\dot{x})$. Then $\mathrm{S}^+ \vdash \forall x : T\phi(\dot{x})$ and so $\mathrm{S}^+ \vdash \forall x\phi$ by CONEC^+. Therefore the rule *from* $Bew_{S'}\,\phi(\dot{x})$ *infer* $\phi(x)$ for ϕ in $\mathcal{L}$ is admissible in S^+ and $\mathrm{S}^+ \vdash \mathrm{RFN}_{\mathcal{L}}(\mathrm{S}')$. □

Contrary to the case with $\mathrm{Out}^-_{\mathcal{L}}$, the addition of CONEC alone does not imply the derivation of new arithmetical theorems.

PROPOSITION 5.4. *Let S be a Σ^0_1-sound theory in the language of $\mathcal{L}_{PA}$ and suppose S^T is the expansion of S to the language $\mathcal{L}_T$. Then* $S^T + \mathrm{GRP}_{\mathcal{L}_T}(S^T) + \mathrm{CONEC}$ *is a conservative extension of S.*

Proof. Define $^* : \mathcal{L}_T \rightarrow \mathcal{L}_{PA}$ to be the interpretation that commutes with all connectives and quantifiers, leaves arithmetical formulae unchanged, and maps $T\phi$ to $Bew_{S^T}\,\phi$. We claim that if ϕ is derivable in $S^T + \mathrm{GRP}_{\mathcal{L}_T}(S^T) + \mathrm{CONEC}$, ϕ^* is a theorem of S. This is proved by induction on the number of applications of CONEC in the derivation. If $S^T + \mathrm{GRP}_{\mathcal{L}_T}(S^T) \vdash \phi$, clearly $S \vdash \phi^*$. On the other hand, if ϕ is obtained by an application of CONEC to $T\phi$, then the induction hypothesis yields $S \vdash Bew_{S^T}\,\phi$ whence, since S is Σ^0_1-sound, $S^T \vdash \phi$, and so $S \vdash \phi^*$ as required. □

That said, with CONS at hand, the global reflection principle does yield new arithmetical theorems.

---

[7] For a detailed presentation of arithmetical reflection principles we refer the reader to Smorynski (1977).

PROPOSITION 5.5. *$S + \mathrm{CONS} + \mathrm{GRP}_{\mathcal{L}_S}(S)$ conservatively extends $S + Consis(S)$.*

Proposition 5.3 shows that over FS5 (and hence over *FS*), the rules CONEC and CONEC^+ are equivalent. As Proposition 5.5 suggests, with *PosFS* as a base theory there is a stark difference between the two rules.

The real interest with global reflection principles is not in their single application but rather in iterating them. As we have seen, $PosFS + \mathrm{GRP}_{\mathcal{L}_T}(PosFS)$ is stably true and hence so is its own global reflection principle, $\tau_1 \equiv \mathrm{GRP}_{\mathcal{L}_T}(PosFS + \mathrm{GRP}_{\mathcal{L}_T}(PosFS))$. But then so is the global reflection principle $\tau_2 \equiv \mathrm{GRP}_{\mathcal{L}_T}(PosFS + \tau_1)$ and so on.

In the following we shall analyze iterated global reflection over two theories: *PosFS* and $PosFS + \mathrm{CONEC}^+$.

DEFINITION 5.6 *Let $PosFS^+$ denote $PosFS + \mathrm{CONEC}^+$. We define, for each ordinal κ, two formulae expressing iterated global reflection over, respectively, PosFS and $PosFS^+$:*

$$\begin{aligned} \tau_\kappa &\equiv \mathrm{GRP}_{\mathcal{L}_T}(PosFS_{<\kappa}). \\ \tau_\kappa^+ &\equiv \mathrm{GRP}_{\mathcal{L}_T}(PosFS^+_{<\kappa}), \end{aligned}$$

where $PosFS_{<\kappa}$ and $PosFS^+_{<\kappa}$ denote the theories $PosFS + \{\tau_\lambda \mid \lambda < \kappa\}$ and $PosFS^+ + \{\tau_\lambda^+ \mid \lambda < \kappa\}$ respectively. We write $PosFS_\kappa$ and $PosFS_\kappa^+$ to abbreviate $PosFS + \tau_\kappa$ and $PosFS^+ + \tau_\kappa^+$.

It should be clear that both $PosFS_\kappa$ and $PosFS_\kappa^+$ are stably true for all κ; they become true at stage $\omega \times (1 + \kappa)$. Moreover, as corollaries of Propositions 5.3 and 5.5 we immediately obtain lower bounds on their proof-theoretic strength in terms of iterated consistency and reflection principles over arithmetic.

DEFINITION 5.7 *Let the theories PA_κ, PA_κ^*, and PA_κ^- denote the expansion of PA by, respectively, κ-times iterated uniform reflection, κ-times iterated Π^0_2 uniform reflection, and κ-times iterated consistency. Explicitly, for each ordinal κ, PA_κ is defined as the theory $PA + \bigcup_{\lambda<\kappa} \mathrm{RFN}_{\mathcal{L}_{PA}}(PA_\lambda)$, PA_κ^* is the theory $PA + \bigcup_{\lambda<\kappa} \mathrm{RFN}_{\Pi^0_2}(PA_\lambda^*)$, and PA_κ^- denotes $PA + \bigcup_{\lambda<\kappa} Consis(PA_\lambda)$.*

PROPOSITION 5.8. *$PA^-_{\omega\times\kappa}$ is a subtheory of $PosFS_{<\kappa}$ and PA_κ is a subtheory of $PosFS^+_{<\kappa}$ for every κ.*

Proof. The second claim, namely that PA_κ is a subtheory of $PosFS^+_{<\kappa}$ is a consequence of iterating Proposition 5.3. The first part is a corollary of Proposition 5.5 with the observation that if R is a theory with only modus ponens as a rule of inference then $S + \mathrm{NEC} + \phi + \mathrm{GRP}_{\mathcal{L}}(R) \vdash \mathrm{GRP}_{\mathcal{L}}(R + \phi)$. □

5.2. Reflecting on positive truth. We seek to determine upper bounds on the arithmetical strength of $PosFS_\kappa$ (and $PosFS_\kappa^+$) for each κ (bounded, say, by Γ_0). This occurs in three steps. We first stratify the construction of $PosFS_\kappa$, dropping the rule co-necessitation and distinguishing applications of necessitation, to obtain a hierarchy of theories P_α for $\alpha < \omega \times (\kappa + 1)$. Then it is proven that co-necessitation is admissible in each P_α, whence we deduce that each theorem of $PosFS_\kappa$ is derivable in P_λ for some $\lambda < \kappa \times (\omega+1)$. Finally, we prove that each level of the hierarchy can be interpreted in the appropriate extension of *PA*, whereby we obtain an embedding of $PosFS_\kappa$ into arithmetic that preserves truth-free formulae.

We begin with the theories $PosFS_\kappa$, that is κ-times iterated global reflection over *PosFS*. Define a transfinite hierarchy of theories P^n_α for $\alpha < \Gamma_0, n < \omega$ as follows. P^0_0 is the theory whose axioms are those of *PosFS*. Note that P_0^0 is not by definition closed under either of the rules NEC or CONEC. For each $\alpha > 0$ and $n < \omega$ we then set

$$P_\alpha^{n+1} = P_\alpha^n + \{T\psi \mid P_\alpha^n \vdash \psi\}$$
$$P_\alpha^0 = P_{<\alpha} + \mathrm{GRP}_{\mathcal{L}_T}(PosFS_{<\alpha}),$$

where $P_{<\alpha}$ is the collection of axioms of P_β^n for each $\beta < \alpha$ and each n.

Notice that if $P_\alpha^n \vdash \phi$ then $P_\alpha^{n+1} \vdash T\phi$, so for each κ, $P_{<\kappa}$ forms a theory closed under NEC. If we establish that P_α^n is also closed under CONEC for every $\alpha < \kappa$ and $n < \omega$, then clearly $PosFS_\kappa$ is a subtheory of $P_{<\kappa+1}$.

In essence, the reduction of the $PosFS_\alpha$ hierarchy to the P_α^n hierarchy will proceed by formalizing the model-theoretic proof that $PosFS_\alpha$ is stably true. In place of the semantic predicate $\mathfrak{M}_\alpha \models \phi$ we will use the formal predicate $S_\alpha \vdash \phi^{\dagger\alpha}$, where S is some suitably chosen arithmetical theory and ${}^{\dagger\alpha}$ is an interpretation from $\mathcal{L}_T$ into $\mathcal{L}_{PA}$. As in the consistency proof for $PosFS_\alpha$, we argue by transfinite induction, showing that all theorems of $PosFS_{<\alpha}$ are stably true, whence deducing that $\mathrm{GRP}_{\mathcal{L}_T}(PosFS_{<\beta})$ is too.

We can now fix the interpretations ${}^{\dagger\alpha}$ and target theory S_α that we shall work with. These will turn out to be the natural choices for carrying out the proof (S_α will not only be the target theory, but also the background theory for the reduction). We cannot use PA_α^- as a metatheory because to prove that P_β^n is closed under CONEC we require a theory that contains at least (certain instances of) Σ_1^0 reflection. However PA_α^* does suffice for playing the role of S_α. The interpretation ${}^{\dagger\alpha}$ is defined so that

a) ${}^{\dagger\alpha}$ commutes with all connectives and quantifiers;
b) $\phi^{\dagger\alpha} = \phi$ if ϕ is in $\mathcal{L}_{PA}$;
c) $(T\psi)^{\dagger\alpha}$ is the formula $\exists\gamma\,\exists n : \omega \times \gamma + n < \alpha \wedge Bew_{P_\alpha^n}\,\psi$.

A crucial part of the proof outlined above is the use of transfinite induction. In formal systems the schema of transfinite induction for Π_n^0 formulae up to the ordinal $\kappa < \Gamma_0$, denoted $\mathrm{TI}_n(\kappa)$, is the schema

$$(\forall\alpha : \forall\beta < \alpha\phi(\beta) \to \phi(\alpha)) \to \forall\alpha < \bar{\kappa} : \phi(\alpha)$$

for each Π_n^0 formula ϕ.

LEMMA 5.9. *Suppose $\omega \times \alpha < \kappa$. Then the following are derivable in PA_κ^**

1. $\forall\beta \le \bar{\alpha}\forall n\forall\phi : P_\beta^n \vdash \phi \to (\forall\gamma \ge \omega \times \beta + n)PA_\gamma^* \vdash \phi^{\dagger\gamma}$.
2. $\forall\beta \le \bar{\alpha}\forall n\forall\psi : P_\beta^n \vdash T\psi \to P_\beta^n \vdash \psi$.
3. $\forall\beta \le \bar{\alpha}\forall\phi : PosFS_\beta^- \vdash \phi \to P_{<\beta+1} \vdash \phi$.

Proof. We proceed by (meta-)transfinite induction on κ arguing within PA_κ^*. Each of 1, 2, and 3 is Π_2^0 and since the schema $\mathrm{TI}_2(\kappa)$ is present in PA_κ^* (Schmerl, 1979) we will utilize formal transfinite induction on $\beta \le \alpha$ and, in the case of 1, 2 with a subsidiary induction on n.

We begin with 1. Fix $\gamma \ge \omega \times \beta + n$ and suppose $P_\beta^n \vdash \phi$. We proceed by induction on n. Suppose $P_\beta^n \vdash \phi$. There are four prevailing cases to consider:

a) $n = 0$ and $P_{<\beta} \vdash \phi$;
b) $n = 0$ and $\phi = \mathrm{GRP}_{\mathcal{L}_T}(PosFS_{<\beta})$;
c) $n = m + 1 > 0$ and ϕ is an axiom of P_β^m;
d) $n = m + 1 > 0$ and $\phi = T\psi$ with $P_\beta^m \vdash \psi$.

Suppose case (a) applies. Then ϕ is an axiom of the theory P_δ^p for some $\delta < \beta$ and $p < \omega$, whence $\gamma > \omega \times \delta + p$ and the induction hypothesis for 1 yields $PA_\gamma^* \vdash \phi^{\dagger\gamma}$ as desired. Case (c) is similar. (b) is a consequence of the main induction hypothesis for 3, which implies $PA_\beta^* \vdash \forall \phi : PosFS_{<\beta}^- \phi \rightarrow P_{<\beta} \vdash \phi$. Finally, to see (d) suppose $P_\beta^m \vdash \psi$ for some $m < n$. Then $PA \vdash Bew_{P_\beta^m}\, \psi$, so $PA_\gamma^* \vdash (T\psi)^{\dagger\gamma}$ for every $\gamma \geq \omega \times \beta + n$. The case that ϕ is not an axiom of P_β^n is standard an hence omitted.

We now address 2. Suppose $P_\beta^n \vdash T\psi$ for some n. Let $\delta = \omega \times \beta + n$. By 1, $PA_\delta^* \vdash (T\psi)^{\dagger\delta}$. Since $\beta \leq \alpha$, $\delta < \kappa$ and $\forall \psi : Bew_{PA_\delta^*} \ulcorner (T\psi)^{\dagger\delta} \urcorner \rightarrow (T\psi)^{\dagger\delta}$ is an axiom, so $P_\beta^n \vdash \psi$ follows by the definition of ${}^{\dagger\delta}$.

Given 2, $P_{<\beta+1}$ forms a theory closed under both NEC and CONEC. Moreover, by definition, $P_{<\beta+1}$ contains as an axiom the global reflection principle for $PosFS_{<\beta}$, whence $PosFS_\beta$ is clearly a subtheory of $P_{<\beta+1}$. □

The decision to include $\mathrm{GRP}_{\mathcal{L}_T}(PosFS_{<\alpha})$ as an axiom of P_α^0 and not $\mathrm{GRP}_{\mathcal{L}_T}(P_{<\alpha})$ is technically motivated. If the reflection principle for $P_{<\alpha}$ was chosen, difficulties would arise in the embedding of $PosFS_{<\alpha}$ into $P_{<\alpha}$: In order to prove that $PosFS_\alpha$ is a subtheory of $P_{<\alpha+1}$, we must establish $P_{<\alpha+1} \vdash Bew_{PosFS_{<\alpha}}\, \phi \rightarrow Bew_{P_{<\alpha}}\, \phi$. To achieve this the induction hypothesis of Lemma 5.9(3) needs to be derivable in $P_{<\alpha+1}$. By our choice of P_α^0, however, this induction hypothesis can be kept external to $P_{<\alpha+1}$.

As a consequence of Lemma 5.9, an upper bound on $PosFS_\kappa$ is obtained.

COROLLARY 5.10. *For each $\kappa < \Gamma_0$, all arithmetical theorems of $PosFS_{<\kappa}$ are derivable in $PA_{\omega\times\kappa}^*$.*

Now we turn our attention to the theories $PosFS_\kappa^+$. We cannot hope to embed $PosFS_{<\kappa}^+$ into $P_{<\lambda}$ for any λ as $P_{<\lambda}$ is not in general closed under CONEC$^+$ (if it were it would derive λ-times iterated uniform reflection and hence would not be directly interpretable in $PA_{\omega\times\lambda}^*$). Therefore the intermediate hierarchy must be altered, as must the target theory.

Suppose a hierarchy of theories Q_α^n is defined in a similar fashion to P_α^n; that is $T\psi$ is an axiom of Q_β^{n+1} whenever $Q_\beta^n \vdash \psi$. Moreover, suppose we fix a similar interpretation, namely $T\psi$ is interpreted in Q_β^{n+1} as provability in Q_β^n. For Q_β^{n+1} to be closed under CONEC$^+$ we must have $Q_\beta^{n+1} \vdash \forall x \phi(x)$ whenever $Q_\beta^{n+1} \vdash \forall x : T\phi(\dot{x})$. However, the presumed interpretation of $Q_\beta^{n+1} \vdash \forall x : T\phi(\dot{x})$ is that $\phi(\bar{n})$ is provable in Q_β^n for every n, so it follows that Q_β^{n+1} must be closed under some form of ω-logic. This is exactly the additional requirement that we place on the construction of Q_β^n.

Given S (a theory or a set of formulae), let $\omega(S)$ denote the set of theorems of S where one use of the ω-rule is permitted. That is, if $S \vdash \phi(\bar{n})$ for every n, then $\forall x \phi$ is in $\omega(S)$.

We can now define the theories Q_α^n:

$$\begin{aligned} Q_0^0 &= P_0. \\ Q_\alpha^{n+1} &= \omega(Q_\alpha^n) + \{T\psi \mid Q_\alpha^n \vdash \psi\}. \\ Q_\alpha^0 &= \omega(Q_{<\alpha}) + \mathrm{GRP}_{\mathcal{L}_T}(PosFS_{<\alpha}^+). \end{aligned}$$

We also introduce a new interpretation ${}^{*\alpha}$ as indicated above. ${}^{*\alpha}$ commutes with all connectives and quantifiers, leaves arithmetical statements untouched, and

$$(T\psi)^{*\alpha} = \exists\gamma\, \exists n : Bew_{Q_\gamma^n}\, \psi \wedge \omega \times \gamma + n < \alpha.$$

By a straightforward induction on κ we obtain the following result.

LEMMA 5.11. *For every κ, $PA_{\kappa+1} \vdash \forall\gamma\,\forall n : \omega \times \gamma + n < \bar{\kappa} \rightarrow Consis(Q^n_\alpha)$.*

LEMMA 5.12. *Let α, κ be such that $\omega \times \alpha < \kappa$. Then the following is derivable in PA_κ.*

1. $\forall\beta \leq \bar{\alpha}\forall n\forall\phi : Q^n_\beta \vdash \phi \rightarrow (\forall\gamma \geq \omega \times \beta + n)PA_{\gamma+1} \vdash \phi^{\dagger\gamma}$.
2. $\forall\beta \leq \bar{\alpha}\forall n\forall\ulcorner\psi(x)\urcorner : Q^n_\beta \vdash \forall x T\psi(\dot{x}) \rightarrow Q^n_\beta \vdash \forall x\psi(x)$.
3. $\forall\beta \leq \bar{\alpha}\forall\phi : PosFS^+_\beta \vdash \phi \rightarrow Q_{<\beta+1} \vdash \phi$.

Proof. The proof proceeds analogously to Lemma 5.9. We will outline the main differences. In establishing 1, the equivalent subcases (a) and (c) must be amended to cover applications of the ω-rule. For case (a) the new reasoning is as follows (case (c) is similar).

a) If $n = 0$, $\phi = \forall x\psi$ and $Q_{<\beta} \vdash \psi(\bar{n})$ for each n, the induction hypothesis implies $\forall x : PA_\gamma \vdash \psi^{*\gamma}(\dot{x})$, whence $PA \vdash \forall x : Bew_{PA_\gamma}\,\psi^{*\gamma}(\dot{x})$ and so, by reflection, $PA_{\gamma+1} \vdash (\forall x\psi)^{*\gamma}$.

The argument for 2 is the same as before: Suppose $Q^n_\beta \vdash \forall x T\psi(\dot{x})$. By 1, $PA_{\omega\times\beta+n} \vdash (T\psi(\bar{p}))^{*\gamma}$ for every p. Since $\omega \times \beta + n < \kappa$, reflection yields $\forall x : Q^m_\gamma \vdash \psi(\dot{x})$ for some γ, n with $\omega \times \gamma + m < \omega \times \beta + n$, and hence $Q^n_\beta \vdash \forall x\psi$ by definition.

Given 1 and 2, 3 is immediate. □

Combining the the previous lemma with the earlier results, we can characterize the strength of iterated global reflection over *PosFS*.

THEOREM 5.13. *Let $\kappa = \omega^\lambda$ where $\lambda > \omega$. Then*

1. *$PosFS^+ + \{\tau^+_\lambda \mid \lambda < \kappa\}$ and PA_κ have the same arithmetical theorems.*
2. *All arithmetical theorems of $PosFS + \{\tau_\lambda \mid \lambda < \kappa\}$ are derivable in PA^*_κ. Moreover, if $\kappa = \omega^\kappa$ then $PosFS + \{\tau_\lambda \mid \lambda < \kappa\}$ and PA^-_κ have the same Π^0_1 arithmetical theorems.*

Proof. By Proposition 5.8, PA^+_κ and PA^-_κ are subtheories of $PosFS^+_{<\kappa}$ and $PosFS_{<\kappa}$ respectively, while Lemmata 5.9 and 5.12 entail $PosFS^+_{<\kappa}$ and $PosFS_{<\kappa}$ are subtheories of PA^+_κ and PA^*_κ since $\omega \times \kappa = \omega^{1+\lambda} = \omega^\lambda = \kappa$. Finally, if $\kappa = \omega^\kappa$ then PA^-_κ and PA^*_κ prove the same Π^0_1 statements (Schmerl, 1979). □

In sum, for weakly closed ordinals κ, the arithmetical strength $PosFS^+ + \{\tau^+_\lambda \mid \lambda < \kappa\}$ is the same as that of PA_κ. So even though uniform reflection can be unproblematically added to $PosFS^+$—which cannot not be consistently done for *FS*—iterated adding of global reflection to $PosFS^+$ is not a significantly stronger process than iterated adding of uniform reflection to *PA*.

From a recursion theoretic point of view, the collection of nearly stable truths are just as complex as the collection of stable truths. Yet when we look at a natural collection of first-order principles governing stable truth ($PosFS^+$), we see that it is significantly weaker than the corresponding natural collection of first-order principles governing nearly stable truth (*FS*). In particular, this highlights the strength of the nearly stably but not stably true principle FS5 (the right-to-left direction). The strength of this principle, which expresses the infinitary closure of the truth predicate, was already emphasized in Sheard (2002, p. 179).

5.3. Hierarchies of stable truth. Strengthening *PosFS* by adding reflection principles is not the only way. The liar sentence L can be used to form sentences that have a similar effect. Let σ_0 be the sentence $\neg TL \wedge \neg T\neg L \wedge T(0 = 0)$. It is not hard to see that the sentence $\exists n T^n\sigma_0$ becomes true for the first time at stage ω and stays nearly stably true forever after. Thus we are led to construct the following hierarchy:

$$\sigma_0 = \neg TL \wedge \neg T\neg L \wedge T(0=0),$$
$$\sigma_{\kappa+1} = \neg TL \wedge \neg T\neg L \wedge T\sigma_\kappa$$
$$\sigma_\lambda = \forall \kappa < \lambda : \exists n T^n \sigma_\kappa \text{ for } Lim(\lambda)$$

Using the techniques of Section §3, the required liar-like sentences σ_α can be found for all $\alpha < \zeta$. Then, as before with the reflection principles τ_κ, the sentence $\exists n T^n \sigma_\kappa$ becomes *stably* true at stage $\omega \times \kappa$.

This means that semantic deficiency can be asserted by axiomatizations of nearly stable and of stable truth without introducing a new nontruthfunctional connective as is done in Field (2008). But the sentences of the σ-hierarchy are not natural candidates for basic truth axioms. We leave it as an open problem whether natural axioms describing stable or nearly stable truth can be found that entail the semantic deficiency of a large collection of paradoxical sentences.

In contrast with Feferman's system *KF*, the theory *FS* only proves finite truth-iterations (Horsten, 2011, p. 109, Proposition 48). In fact, *FS* proves only nearly stable truths that become nearly stably true already before stage ω in the sequence of revision models. Similarly, *PosFS* proves only stable truths that become stably true at some finite stage in the revision process. This is an indication that *FS* and *PosFS* are very far from capturing the spirit of the notions of nearly stable and of stable truth.

In an attempt to find natural candidates for strengthening *FS* and *PosFS*, we may wonder if we can strengthen our axiomatization of the revision-theoretic truths by strengthening the inference rules NEC and CONEC to their infinitary cousins. This would of course enable the resulting systems to prove transfinite truth-iterations.

For the nearly stable truths, this idea does not work:

PROPOSITION 5.14. *The nearly stable truths are not closed under the inference rules*

$$\phi \Rightarrow \forall n : T^n(\phi)$$

and

$$\exists n : T^n(\phi) \Rightarrow \phi.$$

Proof. We know that McGee's sentence γ is nearly stably true. We also know that $\gamma \leftrightarrow \exists n : T^n(\gamma)$ is true everywhere by the fixed point property. So $\forall n : T^n(\gamma)$ cannot be nearly stably true.

$\neg T(L) \wedge \neg T(\neg L)$ is only true at limit stages, so it is not nearly stably true. But $\exists n : T^n(\neg T(L) \wedge \neg T(\neg L))$ is stably true and hence nearly stably true. □

This confirms earlier results by Gupta and Belnap that the truth iteration laws of nearly stable truth are somewhat unnatural (Gupta & Belnap, 1993, p. 221).

For the stable truths, this strategy does meet with some degree of success:

PROPOSITION 5.15. *The stable truths are closed under the inference rule*

$$\phi \Rightarrow \forall n : T^n(\phi).$$

Proof. Straightforward. □

But it also holds that:

PROPOSITION 5.16. *The stable truths are not closed under the inference rule*

$$\exists n : T^n(\phi) \Rightarrow \phi.$$

Proof. First, we note that the revision theory classifies $\gamma, T\gamma, T^2\gamma, \ldots$ as paradoxical. McGee's sentence γ is made true by successor models. But at limit models, γ is always false. This shows that γ is paradoxical, but it also shows that each $T^n\gamma$ is paradoxical.

Second, we see that the sentence $\exists n T^n \gamma$ is a stable truth. ($\exists n T^n \gamma$ becomes a stable truth from stage ω onwards.) $\square$

Thus not only can we obtain new truth principles from the revision theory, but it works also the other way round: we can learn more about the revision theory of truth using axiomatic truth theories.

Question 3 *What is the arithmetical strength of adding to PosFS or PosFS$^+$ the infinitary version of NEC?*

We do not propose *PosFS*$^+$ + NEC$^{<\Gamma_0}$ as a natural axiomatization of stable truth. The reason is that the generalization of the necessitation rule that is involved explicitly mentions transfinite ordinals. Such a rule is not plausibly taken to be a fundamental truth rule. The system *PosFS*$^+$ + NEC$^{<\omega}$ (as in Proposition 5.15), however, is a more natural theory of truth. After all, the natural numbers are already presupposed in the background theory, and they are quantified over in some of the compositional axioms of *FS* and of *PosFS*—for instance in the axioms that describe how the truth predicate commutes with the quantifiers.

§6. Acknowledgments. Versions of this article have presented at the *Truth Be Told* conference in Amsterdam (March 2011), at the *Truth At Work* conference in Paris (June 2011), at the workshop on *Set Theory and Higher-Order Logic* in London (August 2011), and a seminar in München (April 2011). The authors are grateful to the audiences at these conferences, as well as to an anonymous referee, for helpful comments and suggestions. The work of the second author was supported by the Arts & Humanities Research Council UK grant AH/H039791/1. The work of the third author was supported by the Alexander von Humboldt Foundation.

BIBLIOGRAPHY

Beall, J. C., editor. (2007). *Revenge of the Liar. New Essays on the Paradox*. Oxford, UK: Oxford University Press.

Burgess, J. B. (1986). The Truth is never simple. *Journal for Symbolic Logic*, **51**(3), 663–681.

Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, **56**, 1–49.

Feferman, S. (2008). Axioms for determinateness and truth. *Review of Symbolic Logic*, **1**, 204–217.

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

Field, H. (2008). *Saving Truth from Paradox*. New York: Oxford University Press.

Friedman, H., & Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, **33**, 1–21.

Gupta, A., & Belnap, N. (1993). *The Revision Theory of Truth*. Cambridge, MA: MIT Press.

Halbach, V. (1994). A system of complete and consistent truth. *Notre Dame Journal of Formal Logic*, **35**, 311–327.

Halbach, V. (2011). *Axiomatic Theories of Truth*. Cambridge, UK: Cambridge University Press.

Halbach, V., & Horsten, L. (2005). The deflationist's axioms for truth. In Beall, J. C., and Armour-Garb, B., editors. *Deflationism and Paradox*. Oxford, UK: Clarendon Press, pp. 203–217.

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

Horsten, L. (2011). *The Tarskian Turn. Deflationism and Axiomatic Truth*. Cambridge, MA: MIT Press.

Kripke, S. (1975). *Outline of a Theory of Truth*. Reprinted in Martin (1984, pp. 53–81).

Martin, R., editor. (1984). *Recent Essays on Truth and the Liar Paradox*. Oxford, UK: Oxford University Press.

McGee, V. (1985). How truth-like can a predicate be? A negative result. *Journal of Philosophical Logic*, **14**, 399–410.

Schmerl, U. R. (1979). A fine structure generated by reflection formulas over Primitive Recursive Arithmetic. In Boffa, M., van Dalen, D., and McAloon, K., editors. *Logic Colloquium '78*. Amsterdam, The Netherlands: North Holland, pp. 335–350.

Sheard, M. (2002). Truth, provability, and naive criteria. In Halbach, V., and Horsten, L., editors. *Principles of Truth*. Frankfurt, Germany: Hänsel-Hohenhausen, pp. 169–181.

Smorynski, C. (1977). The incompleteness theorems. In Barwise, J., editor. *Handbook of Mathematical Logic*. Amsterdam, The Netherlands: North-Holland, pp. 821–866.

Turner, R. (1990). Logics of truth. *Notre Dame Journal of Formal Logic*, **31**, 308–329.

Welch, P. D. (2001). On Gupta-Belnap revision theories of truth, Kripkean fixed points, and the next stable set. *Bulletin of Symbolic Logic*, **7**, 345–360.

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

Welch, P. D. (2011). Truth, logical validity, and determinateness: A commentary on Field's *"Saving Truth from Paradox." Review of Symbolic Logic*, **4**(3), pp. 348–359.

Welch, P. D. (Forthcoming). *Some Observations on Truth Hierarchies*. Submitted for publication.

LEON HORSTEN
DEPARTMENT OF PHILOSOPHY
UNIVERSITY OF BRISTOL
43 WOODLAND ROAD
BRISTOL BS81UU, UK
E-mail: Leon.Horsten@bristol.ac.uk

GRAHAM LEIGH
FACULTY OF PHILOSOPHY
UNIVERSITY OF OXFORD
10 MERTON STREET
OXFORD, OX14JJ, UK
E-mail: graham.leigh@philosophy.ox.ac.uk

HANNES LEITGEB
FAKULTÄT FÜR PHILOSOPHIE, WISSENSCHAFTSTHEORIE
UND RELIGIONSWISSENSCHAFT
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
GESCHWISTER-SCHOLL-PLATZ 1
D-80539 MÜNCHEN, GERMANY
E-mail: Hannes.Leitgeb@lmu.de

PHILIP WELCH
SCHOOL OF MATHEMATICS
UNIVERSITY OF BRISTOL
HOWARD HOUSE
UNIVERSITY WALK
BRISTOL BS81TW, UK
E-mail: P.Welch@bristol.ac.uk