Inferentialist Deflationism

Jönne Speck*

31 Aug 2010

Introduction

According to the deflationist about truth, the English expression '... is true' (the 'truth predicate') does not stand for a property. To say that 'Snow is white' is true is just saying that snow is white.

However, little agreement has been achieved so far how this 'just' is to be understood. Leon Horsten [Horsten, 2009] has now set out to exploit the resources of another, more definite programme: inferentialism. Horsten argues for *inferentialist deflationism* about truth:[1] there is nothing to truth but a set of inference rules that govern the truth predicate.

Horsten derives his inferentialist deflationism from a specific reading of formal truth theory. In the following, I will argue that this approach fails.

The bulk of the paper consists of sections 1 and 2 in which I examine and evaluate Horsten's argument for inferentialist deflationism. Section 1 identifies its logical and philosophical presuppositions and provides some necessary background. Section 2 challenges Horsten's main premise that the best formal account of truth currently available does not prove universal quantifications into the truth predicate. I argue that a counterexample is found in the theory developed by Hartry Field [Field, 2003, Field, 2007, Field, 2008]. Horsten's treatment of Field's theory I find unsatisfactory (§2.2.1). I develop an alternative response on his behalf: Since Field's theory is of limited expressive

*jspeck@runbox.com

[1] Horsten speaks of 'inferential deflationism'. I hope that my terminology clarifies where the proposal is located in conceptual space.

power it cannot serve the deflationist's purpose, I discuss Field's claim of having advanced a *revenge-immune* solution to the paradoxes (§2.3) but then turn to argue against Horsten's implicit assumption that only *semantically self-sufficient* formal theories are relevant for the deflationist project (§2.3.1). I conclude that Horsten's argument fails.

However, I do not think this makes inferentialist deflationism a lost cause. The last section of the paper sketches an alternative argument for inferentialist deflationism about truth (§3.2) that I consider more promising.

1 Inferentialist Deflationism

1.1 Horsten's Argument

Horsten summarizes his argument as follows.

> '(. . .) it is of critical importance for a philosophical discussion of truth to focus on the best formal truth theory that is currently available.' [Horsten, 2009, p.1]
>
> 'It will be argued that some proof-theoretic variant of Kripke's theory fits this bill. A particular proof-theoretic version of Kripke's theory of truth is taken to be currently our best formal theory of truth.' (ibid.)
>
> 'It will be argued that sound proof-theoretic versions of Kripke's theory of truth do not contain truth axioms at all but consist entirely of inference rules governing the notion of truth. And this is essentially so. This suggests that truth is essentially an inferential notion.' [Horsten, 2009, p. 2]

I infer that Horsten argues for inferentialist deflationism on the basis of three premises:

P1 Philosophy of truth needs be based on the best formal truth theory currently available.

P2 The best such theory is an axiomatization of Kripke's fixed point models.

P3 Necessarily, this theory's truth predicate is not governed by axioms, only by inference rules.

(P1) and (P2) together imply that a philosophical account of truth must be based on an axiomatic variant of Kripke's theory. Since by (P3) this theory provides only rules of inference, and necessarily so, it follows that truth should be explained as an *inferential* notion.

This inferentialism, however, need not yet be deflationist. Inferential *deflationism* is only established by Horsten's final inference [Horsten, 2009, p. 20].

> 'According to PKF there are no unrestricted general principles of truth. This can be explained by the fact that there is no nature or essence of truth to be described by general principles.'

Horsten admits that this inference is not strictly a valid deduction, but an inference to the best explanation [Horsten, 2009, p. 24]. For the sake of the argument I grant him this move.

More problematic I consider the premises (P1) to (P3). Although none of them is trivial, the justification Horsten provides is cursory. Before I can discuss Horsten's argument for inferentialist deflationism, therefore, I need to fill the gaps he leaves.

1.2 Horsten's First Premise: Deflationism and the Best Formal Theory

Horsten assumes that philosophy of truth ought to be based on a formal truth theory, indeed on the *best* such.[2] Two questions arise immediately. Firstly, what makes a formal theory of truth good, and what makes it the best? Secondly, why should a philosopher bother about *formal* theories in the first place? In the end, these are subject matter of mathematical logic, and their relevance for the philosophical understanding of truth is not obvious.

1.2.1 Philosophy and Formal Theories of Truth

At this point, Horsten's project already needs some qualification. In his article, I do not find an argument as to why *any* philosopher must take into account formal results. Instead, Horsten seems to argue for a more modest thesis [Horsten, 2009, p. 16].

> 'Any concrete version of deflationism has to be articulated against the background of a formal theory of truth.'

Only the deflationist, therefore, needs be committed to formal theory. In this case, however, Horsten's first premise would be too weak to establish inferentialist deflationism. If his argument presupposes deflationism, it cannot establish a competitive philosophy of truth. All that Horsten would be able to show is that the deflationist about truth should be inferentialist: she should take truth to be an inferential notion.

Maybe Horsten really seeks to establish merely this conditional thesis [Horsten, 2009, abstract].

> 'In this article, the prospects of deflationism about the concept of truth are investigated.'

[2] Of course, 'Formal' is a notoriously vague term, see e.g. [MacFarlane, 2000]. What Horsten means by a 'formal theory' are theories in formal languages, as they are subject of mathematical logic.

However, at central places Horsten claims to establish more. [Horsten, 2009, p. 2]

> 'The position of inferentialist deflationism that will be developed and defended in this article claims that the insubstantiality of truth consists in the fact that there simply are no absolutely general laws or principles of truth. Truth is a property without a nature or essence. The content of this property is given by its inferential properties.'

There is no conditional or subjunctive tone in this statement: Horsten argues for inferentialist deflationism as an account of truth.

Nonetheless, since he does not argue why *any* philosopher ought to base her account of truth on formal theory, I need to conclude that Horsten in fact makes a weaker assumption. His first premise applies only to deflationism.

1.2.2 Deflationism and Formal Truth Theory

Why, now, should the deflationist bother about formal theory? Horsten simply takes this as given. Fair enough. Most deflationists at least in recent years do worry a lot about formal theories of truth [Halbach and Horsten, 2003, Beall and Armour-Garb, 2005]. In the following, I assume with Horsten that the deflationist needs a formal theory of truth to explain truth talk. But why need it be the best?

1.2.3 Being the Best

Horsten's assumption that the deflationist needs the best formal theory emerges from his discussion of Paul Horwich's work [Horsten, 2009, p. 16].

> (...) many of Horwich's difficulties arose from the fact that he based his minimalist theory on the disquotational theory of truth, which was already superseded by the compositional theory of truth. So let us see what is the best formal theory of truth available today.

Horsten sums up Horwich's *minimal theory of truth* [Horwich, 1998, Horwich, 1992] in two statements [Horsten, 2009, p. 3]. First, the truth-predicate plays no role in scientific reasoning except that it allows first order theories to quantify over propositions and thereby express infinite conjunctions. Second, the meaning of the truth-predicate is exhausted by the T-schema.

Horsten rightly emphasizes that the latter thesis needs qualification. If for any sentence ϕ, ϕ is true if and only if ϕ, then especially for some liar sentence λ, λ is true if and only if it is not the case that λ is true. Contradiction. The paradoxes can be suppressed if the T-schema is restricted to sentences that do not contain the truth predicate themselves. To fix matters, I follow Horsten and focus on a given language of arithmetic $\mathcal{L}_a$ that is extended by a predicate 'T' to the language $\mathcal{L}_{at}$. If to a classical theory of arithmetic in this language is now added an axiom[3]

$$T\ulcorner\phi\urcorner \leftrightarrow \phi$$

for any $\mathcal{L}_a$-sentence ϕ, a formal theory is obtained that is provably consistent[4]. This theory is called the *disquotational theory* ('DT').

Horsten argues that Horwich's variant of deflationism fails because DT is not good enough a formal theory. DT cannot do what Horwich claims it to do: it does not exhaust the meaning of the truth-predicate of ordinay discourse.

Compositionality Intuitively, if I say that 'Snow is white and grass is green' is true then I am committed to accept also that 'Snow is white' is true and 'Grass is green' is true, and vice versa. Let the degree of this commitment be called the 'compositional' intuition. DT, now, does not prove that the truth predicate commutes with the connectives.

Surely, for any arithmetical sentences ϕ and ψ DT proves $T\ulcorner\phi\urcorner \dot{\vee} \ulcorner\psi\urcorner \leftrightarrow T\ulcorner\phi\urcorner \vee T\ulcorner\psi\urcorner$. However, DT does not prove the universally quantified [5]

$$\forall x \forall y(Sent_a \rightarrow (Tx \dot{\vee} y \leftrightarrow Tx \vee Ty)) \tag{1}$$

Although the theory proves every single instance, it does not tell us the general fact. For Horsten this is reason enough to dismiss DT [Horsten, 2009, p. 6]; it cannot fulfil the role the deflationist needs her formal theory for, it cannot capture ordinary truth talk.

[3] Where $\ulcorner\phi\urcorner$ denotes the Gödel-number of ϕ; the base theory is assumed to provide the machinery to name any sentence ϕ by a number $\ulcorner\phi\urcorner$, and generally to contain its own syntax. In the present arithmetical setting, this is done by representation of some Gödelization function, I adopt the terminology from [Feferman, 1991, §2.1].

[4] Indeed, it is *conservative* over standard arithmetic [Halbach, 1996, ch. 9].

[5] This fact can be traced back to Tarski's seminal *Wahrheitsbegriff* [Tarski, 1956, p. 256, theorem III]. Interestingly, it also holds for the extension of DT examined in [Halbach, 2009] that is proof-theoretically as strong as the theory KF considered in below (p. 10).

Horwich's deflationism fails because it explains the meaning of the truth predicate in terms of a theory that does not prove the universally quantified principles of compositionality. From this, Horsten infers the general methodological principle that makes up his first premise. It can now be phrased more clearly:

P1′ Deflationists need to explain the meaning of the truth predicate by the formal truth theory that proves the most universally quantified principles of truth.

A better choice for Horwich would have been the *compositional theory* ('TC') [Horsten, 2009, p. 5]. It trivially proves (1) as well as compositional principles for the other connectives and the quantifiers since these are taken as axioms, together with all atomic instances of the T-schema[6].

Iteration TC is a good theory, but still not good enough, or so Horsten argues [Horsten, 2009, p. 16f]. The truth predicate of ordinary discourse is self-reflexive. If I say that 'Snow is white' is true then I'm committed to accept that "Snow is white' is true' is true, and vice versa. The degree of this commitment I call the 'iterative intuition'.

TC, however, does not prove that the truth predicate can be iterated:

$$\forall x(\dot{Sent}_a(x) \rightarrow (Tx \rightarrow T\dot{T}x)) \tag{2}$$

. This is easily seen: in the simple model just sketched, only arithmetical sentences are in the extension of 'T' but no sentence $T\dot{T}x$.

For this reason, Horsten dismisses TC as well. Instead, he favours Kripke's theory of truth [Kripke, 1975], or rather an axiomatization of it, as I will explain in the next section.

1.3 The Second Premise: Kripke's Models and their Axiomatization

I now turn to Horsten's second premise:

P2 The best formal truth theory is an axiomatization of Kripke's fixed point models.

Explaining it requires me to go into some technical details. Since these, however, provide also the background of my subsequent discussion, I beg the reader to bear with me.

[6] And, of course, a number theory at least as strong as Q. For a thorough investigation into TC (there called 'T(PA)'), consult [Halbach, 1996, §8].

1.3.1 Kripke's Fixed Point Models and their Theory

Kripke takes a model-theoretic view point: He asks how the extended language $\mathcal{L}_{at}$ is to be *interpreted*[7]. Clearly, the arithmetical sentences are interpreted in the standard model. Furthermore, on this interpretation every such sentence is either true or false: the resulting valuation function m maps every sentence either to 0 or to 1.

For the truth predicate 'T', however, the classical value space is now extended by a third value $\mathfrak{u}$, that may be read 'undefined'[8]. The resulting non-classical value-space $\mathcal{S} = \langle \{0, \mathfrak{u}, 1\}, \leqslant_{\mathcal{S}} \rangle$ makes up a partial ordering ('po') every subset of which with an upper bound in $\{0, u, 1\}$ also has a least upper bound in $\mathcal{S}$ (a 'bound complete' or 'coherent complete' po, henceforth 'ccpo').

The according valuation functions v make up a ccpo $\mathcal{V}$, too, if one defines an order $\leqslant_{\mathcal{V}}$ such that $v \leqslant_{\mathcal{V}} v'$ iff for every $\mathcal{L}_{at}$-sentence ϕ, $v(\phi) \leqslant_{\mathcal{S}} v'(\phi)$[9] [Visser, 2004, lemma 7]. Any operator on $\mathcal{V}$ that preserves the ordering (is *monotone*) has *fixed points*, indeed a ccpo of them.[10]

One way to obtain a monotone operator K is to use the strong Kleene scheme for complex $\mathcal{L}_{at}$-sentences[11].

K1 $K(v)(\phi) = m(\phi)$ for $\phi \in Sent_a$

K2 $K(v)(T^{\ulcorner}\psi^{\urcorner}) = v(\psi)$

K3 $K(v)(\neg\psi) = 1 - K(v)(\psi)$

K4 $K(v)(\psi \vee \chi) = max\{K(v)(\psi), K(v)(\chi)\}$

K5 $K(v)(\forall x\psi) = min\{K(v)(\psi(t/x)) | t \text{ closed term}\}$[12]

[7] There are various ways of presenting Kripke's work. The variant in the main text is meant not only to provide the necessary results but also to prepare the ground for later discussion.

[8] At this point, an exegetical problem arises: Kripke explicitly rejects the interpretation of $\mathfrak{u}$ as a third value – instead, it symbolizes the absence of a truth value [Kripke, 1975, fn 18]. Accordingly, his construction may be better presented by means of *partial* valuation functions. However, I omit this complication, in accordance with most of the contemporary literature.

[9] In the following I omit the subscripts whenever the context determines the ordering.

[10] A corollary of what has become known as the Knaster-Tarski theorem, [Tarski, 1955],[Fitting, 1986, 2.2],[Visser, 2004, 15].

[11] Once we extend our language by the other connectives, corresponding clauses can be added straightforwardly.

[12] Since we deal with arithmetic, it can be assumed that every object has a name in the language.

Notice that by (K1) the sentences that do not contain 'T' keep their values, as interpreted in the standard model. K is easily seen to be monotone[13]. Hence, there is a ccpo of *fixed point* valuations v_f such that $K(v_f) = v_f$. Especially there is a fixed point v_{f0} such that for any ϕ, if $v_{f0}(\phi) = 1(0)$ then for any fixed point $v_f(\phi) = 1(0)$. This *minimal* fixed point valuation has attracted the most attention in the literature and will also be the focus of the present discussion. For the sake of readability, though, I frequently drop the index 0.

From a model-theoretic point of view, the minimal fixed point valuation provides a truth theory: the set of sentences ϕ such that $v_{f0}(\phi) = 1$. It is to this that Horsten refers by 'Kripke's theory', and I follow him. Kripke's theory has a highly desirable feature: in any context ϕ can be replaced by $T^{\ulcorner}\phi^{\urcorner}$, and vice versa.

> The reason is that any sentence ψ that contains ϕ as a subsentence has the same fixed point value as that sentence which results from ψ by replacing one or more occurrences of ϕ by $T^{\ulcorner}\phi^{\urcorner}$ ('$\psi(T^{\ulcorner}\phi^{\urcorner}/\phi)$'). This is shown easily by an induction on the complexity of ψ. If $\psi = \phi$ then the claim follows directly from (K2). Now assume that the complexity of ψ is $n+1$, and for any sentence χ of complexity $\leqslant n$, χ has the same value as $\chi(T^{\ulcorner}\phi^{\urcorner}/\phi)$. If the complexity of ψ is $n+1$ the claim is shown for each possible logical form of ψ separately. If $\psi = \neg\chi$ for some sentence χ then $v_f(\psi(T^{\ulcorner}\phi^{\urcorner}/\phi)) = 1 - v_f(\chi(T^{\ulcorner}\phi^{\urcorner}/\phi))$. By induction assumption, however, $v_f(\chi(T^{\ulcorner}\phi^{\urcorner}/\phi)) = v_f(\chi)$, hence $v_f(\psi(T^{\ulcorner}\phi^{\urcorner}/\phi)) = 1 - v_f(\chi) = v_f(\psi)$. The other cases ($\psi = \chi \vee \chi', \psi = \forall x\chi$) follow in the same manner.
>
> The proof of the converse ($v_f(\psi) = v_f(\psi(\phi/T^{\ulcorner}\phi^{\urcorner}))$) is exactly analogous.

Following Field [Field, 2008, p. 64], I call this feature the *intersubstitutivity* of 'T'. It must not be conflated with the T-schema. On one hand, intersubstitutivity does not suffice to validate $T^{\ulcorner}\phi^{\urcorner} \leftrightarrow \phi$[14]. Even in the fixed points, namely, there remain sentences of value $\mathfrak{u}$. And if ϕ is such a 'gappy' sentence, $T^{\ulcorner}\phi^{\urcorner}$ is, too. Consequently, there are instances of the T-schema that do not have value 1 in any fixed point.

On the other hand, different from the T-schema, intersubstitutivity ensures both the compositionality and the iteration of truth. In fact, for every sentence that has a classical value in the minimal fixed point the principles of compositionality and iteration also hold in the form of object-linguistic conditionals. Especially, Kripke's theory contains (1) and (2) from above. Thus, it seems *better* than both DT and TC.

[13] I omit the proof for the sake of concision.

[14] As usual, this biconditional is the conjunction of material conditionals, which again are defined in terms of $\neg$ and $\vee$.

Horsten disagrees. He rejects Kripke's theory because it is defined in a meta-theory. Since deflationists aim for an account of the *real* truth predicate, they need formal theories that can be applied to ordinary discourse. Kripke's theory, however, does not fit this bill since [Horsten, 2009, p. 17]

> (...) we do not have a metalanguage for English.

Notice that by this reasoning, Horsten implicitly rejects any *semantical* truth theory, i.e. any theory which is obtained by meta-theoretical means. Axiomatic theories, in contrast, go without a meta-theory, or so Horsten assumes. Therefore, an axiomatization of Kripke's theory could serve the deflationist's purpose. In fact, Horsten has in mind a specific axiomatization of Kripke's fixed point models: the theory PKF as developed in his and Volker Halbach's [Halbach and Horsten, 2006], [Horsten, 2009, p. 19].

1.3.2 Axiomatizing Kripke's Theory (I): KF and KFS

However, PKF is not the first attempt to transpose Kripke's truth theory into an axiomatic setting. In fact, it is based on the earliest such axiomatization, formulated by Solomon Feferman[15] and therefore known as 'Kripke-Feferman' or 'KF'[16]. In a sense soon to be specified, PKF translates KF into the Kleene logic of Kripke's models.

Horsten omits this background. This leaves his claim that PKF axiomatizes Kripke's theory unfortunately obscure. Therefore, let me instead tell the whole story[17].

KF Clearly, Kripke's theory cannot be recursively axiomatized in the strict sense of the word, since it contains the theory of the standard model[18]. No formal system can be complete with respect to the minimal fixed point models. Recall, however, that in the fixed point models, 'T' applies to just those sentences of value 1, and its negation to those of value 0. What can be done instead, therefore, is to define axioms for 'T' that correspond to the definitional clauses of K. In this respect the resulting theory KF[19] captures the truth predicate of the minimal fixed point model. Its axioms are:

[15] He presented it at the joint ASL-APL meeting in 1983. In print it appeared first in [Reinhardt, 1986, p. 231f], then in [Cantini, 1989, p. 101] and finally in Feferman's own [Feferman, 1991, §3.2], there under the title 'Ref(PA)'. For a helpful development of KF from the Kripke construction consult [Halbach, 1996, §24f].

[16] The acronym 'PKF' is short for 'Partial Kripke-Feferman'.

[17] As Horsten does, too, in his forthcoming [Horsten, ta, §9.2]

[18] In fact, the set of sentences such that $v_{f0}(\phi) = 1$ is Π^1_1 [Burgess, 1986, §6.1].

[19] Just as DT and TC, KF includes classical first order arithmetic.

KF1a $\forall x, y(Cl\dot{T}erm(x) \wedge ClTerm(y) \rightarrow (Tx \doteq y \leftrightarrow Vl(x) = Vl(y)))$[20]

KF1b $\forall x \forall y(Cl\dot{T}erm(x) \wedge ClTerm(y) \rightarrow (T\dot{\neg}x \doteq y \leftrightarrow Vl(x) \neq Vl(y)))$

KF2a $\forall x(Cl\dot{T}erm(x) \rightarrow (T\dot{T}(x) \leftrightarrow T\,Vl(x)))$

KF2b $\forall x(Cl\dot{T}erm(x) \rightarrow (T\dot{\neg}\dot{T}(x) \leftrightarrow (T\dot{\neg}\,Vl(x) \vee \neg Se\dot{n}t_{at}(Vl(x)))))$

KF2c $\forall x(Tx \rightarrow Se\dot{n}t_{at}(x))$

KF3 $\forall x(Se\dot{n}t_{at}(x) \rightarrow (T\dot{\neg}\dot{\neg}(x) \leftrightarrow Tx))$

KF4a $\forall x \forall y(Se\dot{n}t_{at}(x) \wedge Sent_{at}(y) \rightarrow (T(x \dot{\vee} y) \leftrightarrow Tx \vee Ty))$

KF4b $\forall x \forall y(Se\dot{n}t_{at}(x) \wedge Sent_{at}(y) \rightarrow (T(\dot{\neg}(x \dot{\vee} y)) \leftrightarrow T\dot{\neg}x \wedge T\dot{\neg}y))$

KF5a $\forall x \forall y(Free(x) \wedge Var(y) \rightarrow (T\dot{\forall}(x, y) \leftrightarrow \forall z(TSubst(\ulcorner z \urcorner, y, x))))$

KF5b $\forall x \forall y(Free(x) \wedge Var(y) \rightarrow (T\dot{\neg}\dot{\forall}(x, y) \leftrightarrow \neg\forall z \neg(TSubst(\ulcorner z \urcorner, y, x))))$

Kripke's truth predicate separates the set of sentences into two exclusive parts. No sentence can be both true and false: in this sense Kripke's theory is *consistent*. Usually, therefore, a final axiom is added.

Cons $\forall x(Se\dot{n}t_{at}(x) \rightarrow \neg(Tx \wedge T\dot{\neg}(x)))$

KF is sound with respect to the classical closure of the minimal fixed point valuation[21]. More important for the present study is a corollary: for any ϕ, if KF$\vdash T\ulcorner\phi\urcorner$ then for every α, $v_{f,\alpha}(\phi) = 1$.

The axioms KF4 and KF5 imply that the truth predicate commutes with $\vee$ and $\forall$. The axiom (*Cons*) allows KF to prove also the compositionality of $\rightarrow$ [Halbach and Horsten, 2006, §2]. Thus, KF includes the theory TC from §1.2.3.

[20] I deviate slightly from Feferman's notation. For one, I distinguish the object-linguistic predicates '$Cl\dot{T}erm$' and '$Se\dot{n}t$' from the meta-linguistic '*ClTerm*' and '*Sent*' (no dots). Secondly, I adopt the expression $Vl(t)$ for the value of the term t [Halbach and Horsten, 2006, p. 679f].

[21] That is, a valuation $v_{c0} : Se\dot{n}t_{at} \mapsto \{0, 1\}$ such that

$$v_{c0} = \begin{cases} 1 & \text{iff } v_{f0}(\phi) = 1 \\ 0 & \text{otherwise} \end{cases}$$

. I omit the easy but lengthy proof.

KF is an axiomatic theory. It therefore avoids the need of a meta-theory that Horsten complained of above (§ 1.3.1). Furthermore, it inherits some desirable features of Kripke's theory. In particular, by axiom (KF2a) KF contains the principle of self-reflexivity (1). It therefore outruns the compositional theory and makes up the *best* axiomatic truth theory we have considered so far. So why does Horsten not settle with it as the best formal truth theory currently available, on whose basis deflationism should be developed?

Since Horsten does not mention KF in his paper, I need to look at his reasoning at other places [Halbach and Horsten, 2003, pp. 28f], [Halbach and Horsten, 2005, pp. 209f], [Horsten, ta, § 9.3]. There, he argues that KF fails to be *philosophically* sound because it declares itself untrue: It contains sentences ϕ of which it also proves $\neg T\ulcorner\phi\urcorner$.

> Consider the liar sentence λ.
>
> | | |
> |---|---|
> | 1. $T\ulcorner\lambda\urcorner$ | assumption |
> | 2. $T\dot{\neg}\dot{T}(\ulcorner\lambda\urcorner) \leftrightarrow (T\dot{\neg} Vl(\ulcorner\lambda\urcorner) \vee \neg \dot{Sent}_{at}(Vl(\ulcorner\lambda\urcorner)))$ | KF5b, arithmetic |
> | 3. $T\ulcorner\lambda\urcorner \rightarrow T\dot{\neg}(\ulcorner\lambda\urcorner)$ | 2, logic, arithmetic |
> | 4. $T\dot{\neg}(\ulcorner\lambda\urcorner)$ | 1, 3 |
> | 5. $T\ulcorner\lambda\urcorner \rightarrow \neg T\dot{\neg}(\ulcorner\lambda\urcorner)$ | *Cons*, logic, arithmetic |
> | 6. $\neg T\dot{\neg}(\ulcorner\lambda\urcorner)$ | 1, 5 |
> | 7. $\neg T\ulcorner\lambda\urcorner$ | 1,4,6, logic |
> | 8. $\neg T\ulcorner\lambda\urcorner \leftrightarrow \lambda$ | arithmetic |
> | 9. λ | 7,8, logic |
>
> KF thus proves a sentence that it has just proved untrue (step 7). Notice that the proof requires the axiom *Cons*.

The reason for this odd behaviour is that KF attempts to capture Kripke's non-classical truth theory within a classical framework. Already Reinhardt complained for similar reasons that KF is philosophically unsatisfactory [Reinhardt, 1986, pp. 242f].

KFS He proposed instead to focus on its *significant* part, the sentences that it proves true or false. The corresponding theory KFS[22] is the set of sentences ϕ such that KF proves $T\ulcorner\phi\urcorner$. On one hand, KFS inherits the *intersubstitutivity* property of Kripke's fixed point models.

[22] Halbach and Horsten prefer to call it 'IKF', the *inner logic* of KF [Halbach and Horsten, 2006, p. 683]

The reason is that KF proves $T^{\ulcorner}\psi^{\urcorner}$ just in case it proves $T^{\ulcorner}\psi(T^{\urcorner}\phi^{\urcorner}/\phi)$, for any ϕ subsentence of ψ. This is shown by induction on the complexity of ψ. For atomic ψ, the claim follows easily from KF2a. The induction step requires to distinguish between the different logical forms of ψ, The case for $\psi = \forall x\xi$, for example, follows from KF5.

On the other hand, KFS is not closed under classical implication any longer. For example, it does not contain the classical tautology $\lambda \rightarrow \lambda$, respectively $\lambda \vee \neg\lambda$. It is just this *paracomplete* character of KFS [23] that safes it from KF's undesirable self-refuting character.

Nonetheless, even KFS is not Horsten's preferred theory. Although it is defined by proof-theoretic means it is no axiomatic theory in a strict sense[24]. Its definition ($KFS = \{\phi | KF \vdash T^{\ulcorner}\phi^{\urcorner}\}$) relies on a theory that has been found philosophically unsound, and so far no independent axiomatization of KFS has been found[25].

Neither KF nor KFS therefore transfer the desirable features of Kripke's theory into an axiomatic framework that could serve the deflationist. This, however, is what Horsten claims he and Halbach have achieved by their *PKF* [Halbach and Horsten, 2006, §4].

1.3.3 Axiomatizing Kripke's Theory (II): PKF

Logic Every axiomatic theory considered so far included a system of logic. DT, TC and KF were all based on classical logic. Not so, however, PKF. Just like KFS, it is not closed under classical implication. Since PKF now is meant to be a self-contained axiomatization, however, its logical basis need be made explicit.

The Strong Kleene valuation of Kripke's models is axiomatized in various ways [Stephen Blamey, 2002, Kremer, 1988] Halbach and Horsten use a sequent variant due to Dana Scott [Scott, 1975, §3]. In his (2009) again, Horsten uses a natural deduction calculus, the only rule of which that he specifies is a weakened $\rightarrow$ introduction rule.

[23] A paracomplete account is any that rejects the law of excluded middle (in analogy to the *para-consistent* rejection of non-contradiction). The term is due to JC Beall but has recently been used extensively in [Field, 2006, Field, 2008]. Notice that this approach does not collapse into intuitionism since the so called *Curry* paradox occurs already in minimal logic, see e.g. [Gupta and Belnap, 1993, p.14].

[24] This argument is found only in his forthcoming [Horsten, ta], section 9.3

[25] Except, of course, by Craig's notorious roundabout method.

$$\overset{w}{\rightarrow}\text{I},0\ \dfrac{\begin{array}{c}[\phi^0]\\ \mathcal{D}\\ \psi\end{array} \qquad T^{\ulcorner}\phi^{\urcorner} \vee T\dot{\neg}^{\ulcorner}\phi^{\urcorner}}{\phi \rightarrow \psi}$$

Due to this restricted form of conditional proof, PKF contains only conditionals with truth-determinate antecedents.

Besides, PKF is supposed to contain '(...) the usual introduction and elimination rules (...)' [Horsten, 2009, p. 18]. I suppose he means the classical rules, and therefore work with PKF as if it was based on a variant of the calculus **Nc** from [Troelstra and Schwichtenberg, 1996, §2.1] where $\rightarrow$I is replaced by $\overset{w}{\rightarrow}$I. Unfortunately, Horsten misses to give a proof of this system to be equivalent to Scott's sequent calculus.

Arithmetic As with KF, the base theory of PKF is Dedekind Peano Arithmetic ('PA'). However, PKF does not contain the infinitely many induction axioms but an induction rule:

$$\text{IND},0,1\ \dfrac{\begin{array}{c}[\phi(x)]^1\\ \mathcal{D}\\ \phi(x+1)\end{array} \qquad [\phi(\overline{0})]^0}{\forall x \phi(x)}$$

Thus, ϕ may contain the new predicate 'T'.

Truth Different from the theories considered so far, PKF does not contain axioms for 'T'. There's a simple reason for it [Horsten, 2009, p. 19]. It has been crucial to Kripke's construction that the paradoxical sentences do not receive a classical truth value (1 or 0) but $\mathfrak{u}$. This allowed ϕ and $T^{\ulcorner}\phi^{\urcorner}$ always to have the same value. In consequence, however, for some ϕ, $T^{\ulcorner}\phi^{\urcorner}$ lacks a classical value, too. Therefore, no universal quantification into the predicate 'T' can be true in the fixed point models (see (K5) from p. 8 above) as is required of an axiom.

Instead, the behaviour of 'T' can be described by rules of inference. Also in strong Kleene logic, namely, an inference $\dfrac{\phi}{\psi}$ fails to be valid just in case that the value of ϕ is greater than that of ψ. Thus, even if for some ϕ, $v_f(\phi) = \mathfrak{u}$ such that $v_f(T^{\ulcorner}\phi^{\urcorner}) = \mathfrak{u}$, too,

$$\dfrac{\neg T^{\ulcorner}\phi^{\urcorner}}{T\dot{\neg}^{\ulcorner}\phi^{\urcorner}}$$

still is valid.

Notice that due to the weakened introduction rule for $\rightarrow$ these rules do not give immediate rise to corresponding axioms as they would in a classical context.

Thus, PKF contains no axioms but *rules* of truth[26].

$$\text{PKF1}\ \frac{Vl(x)=Vl(y) \qquad \dot{ClTerm}(x) \qquad \dot{ClTerm}(y)}{T\dot{x=}y}$$

$$\text{PKF2a}\ \frac{T(Vl(x)) \qquad \dot{ClTerm}(x)}{T(\dot{T}(x)) \qquad \dot{ClTerm}(x)} \qquad \frac{T(x)}{\dot{Sent}(x)}\ \text{PKF2b}$$

$$\text{PKF3}\ \frac{\neg T(x) \qquad \dot{Sent}_{at}(x)}{T\dot{\neg}x} \qquad \frac{T(x) \vee T(y) \qquad \dot{Sent}_{at}(x) \qquad Sent(y)}{T x \dot{\vee} y \qquad \dot{Sent}_{at}(x) \qquad Sent(y)}\ \text{PKF4}$$

$$\text{PKF5}\ \frac{\forall z T\dot{Subst}(\ulcorner z \urcorner, y, x) \qquad Free_{at}(x) \qquad Var_{at}(y)}{T\dot{\forall}(x,y)}$$

PKF is a sub-theory of KFS and therefore sound with respect to every fixed point model[27]. Moreover, although incomplete of necessity, it captures the desirable features of Kripke's truth theory in a proof-theoretic setting. Most fundamentally, it contains any sentence ϕ just in case that it contains $T\ulcorner\phi\urcorner$, too [Halbach and Horsten, 2006, theorem 22]. Horsten now argues that PKF is the formal theory that deflationism should be based on.

1.3.4 Why Horsten takes PKF to be the Best Formal Truth Theory

The formal theory traditionally favoured by deflationists, the theory DT from § 1.2.3 proved too weak since it does not contain the principles of compositionality. TC, therefore, included these principles as axioms. However, Horsten argued that TC is still not strong enough because it does not prove the universally quantified principle of iteration (10). The deflationist should therefore use a different theory for her account of ordinary truth talk.

---

[26] $\frac{\Gamma}{\Delta}$ is short for the two rules $\frac{\Gamma}{\Delta}$ and $\frac{\Delta}{\Gamma}$.

[27] That for any ϕ, PKF$\vdash_{sk} \phi$ only if KF$\vdash T\ulcorner\phi\urcorner$ is shown by induction on the length of the proof in PKF. The only non-trivial bit is to show that for any PKF rule $\frac{\Gamma(\vec{x})}{\Delta(\vec{x})}$ KF proves $\forall\vec{x}(\Gamma(\vec{x}) \rightarrow \Delta(\vec{x}))$. Compare Halbach and Horsten's proof of theorem 27 [Halbach and Horsten, 2006], for the sequent calculus variant of PKF.

PKF, now, seems stronger than TC. In fact, Horsten claims, TC is '(...) but a small fragment of PKF' [Horsten, 2009, p. 19].

I take him to refer to a result of his and Halbach's (2006). As theorem 38, they show that PKF proves the compositionality as well as the iteration principles for *ramified truth* up to an ordinal below ω^ω [Halbach and Horsten, 2006, p. 705]. This means, they consider extensions of $\mathcal{L}_{at}$ by predicates

$$T_0, T_1, \ldots, T_\alpha, \ldots, T_\omega, \ldots$$

where $T_0\ulcorner\phi\urcorner$ is well formed only if ϕ is an arithmetical sentence, T_1 applies to $\mathcal{L}_{at0}$-sentences and so on. At limit ordinals, T_λ applies to the union of all $Sent^\alpha_{at}$, for $\alpha < \lambda$. Then, they prove for any $\alpha < \omega^\omega$, that $T_{\alpha+1}$ commutes with the connectives, as universal quantifications over the $\mathcal{L}^\alpha_{at}$-sentences.

> The proof's lemma is that for any α [Halbach and Horsten, 2006, 36],
>
> $$\text{PKF} \vdash (\forall x(Sent^\alpha_{at}(x) \rightarrow (Tx \vee \neg Tx)))$$
>
> . Since this claim involves both a meta- and an object-linguistic universal quantification it is proved by a (transifinite) induction on α. It includes side inductions on the complexity of the sentences 'x' ranges over (Notice that PKF only provides a rule of induction only for natural number exponents of ω, wherefore the upper bound ω^ω). Since now for any $\alpha < \omega^\omega$, the $\mathcal{L}_\alpha$-fragment of PKF is closed under *classical* logic, the PKF-rules amount to universally quantified biconditionals, e.g. PKF4 to
>
> $$\forall x \forall y(Sent^\alpha_{at}(x) \wedge Sent^\alpha_{at}(y) \rightarrow (T_{\alpha+1}(x \dot{\vee} y) \leftrightarrow (Tx \vee Ty)))$$
>
> again for any $\alpha < \omega^\omega$.

In this hierarchy, TC makes up the special case of quantification over 'T'-free, that is purely arithmetical sentences. PKF thus indeed is much stronger than TC.

The same theorem also shows that PKF proves the principles of iteration, not only for $\mathcal{L}_a$-sentences, but for any level up to ω^ω.

> For any $\alpha < \omega^\omega$, PKF proves
>
> $$\forall \beta < \alpha \forall x(Sent_\beta(x) \rightarrow (T_\alpha x \rightarrow T_\beta Sent_\alpha(x))$$

It is on this basis that finally, it becomes clear why Horsten takes PKF to be the best formal truth theory.

P2′ The theory that proves the most universally quantified principles of truth is PKF.

In the subsequent section I will explain which consequences Horsten draws from this for the deflationist project.

1.4 The Inferential Character of PKF

Since Horsten takes PKF to be the best truth theory available today, his third premise becomes:

P3′ Necessarily, the truth predicate of PKF is not governed by axioms but merely by inference rules.

Horsten specifies this claim as follows [Horsten, 2009, p. 19]. First,

> 'PKF proves no unrestricted generalities about truth; for example, it does not provide a proof of any sentence of the form
>
> $$\forall \phi \in L_T : T(...\phi...) \rightarrow T(...\phi...)$$
>
> $(\dots)$'

L_T is Horsten's arithmetical language extended by the untyped truth predicate 'T', it corresponds to my $\mathcal{L}_{at}$. Thus, an 'unrestricted generality about truth' in Horsten's sense is the KF axiom 3, or 4 (section 1.3.2). Since for $\mathcal{L}_{at}$-closed terms s and t, $s \dot{=} t$ is an $\mathcal{L}_{at}$-sentence, he would probably count its other axioms as *generalities*, too. Not so, however, PKF's theorems $\forall x(\dot{Sent}^{\alpha}_{at}(x) \rightarrow (T_{\alpha+1}\dot{\neg}x \leftrightarrow (\neg T_{\alpha+1}x)))$. Even for very large α, $Sent^{\alpha}_{at}$ still is a proper subset of $Sent_{at}$.

Notice that the generality need not be a conditional, Horsten's schema is merely an example. KF's axiom *Cons* is a generality of a different logical form.

Secondly, Horsten writes

> '$(\dots)$ PKF contains lots of unrestricted rules of inference concerning truth.'

and gives an example

> '$T\neg\phi \Rightarrow \neg T\phi$'

by which he must mean the upwards direction of PKF3. In his article, Horsten generally favours readability over technical details, and rightly so. Here, however, his notation obscures what he means by 'unrestricted rules'. My way of putting it[28]

$$\frac{T\dot{\neg}x \qquad Sent_{at}(x)}{\neg Tx}$$

it clarifies what Horsten must have in mind. Namely, it allows to apply the same criterion as before. A rule is unrestricted if its minor premise ranges over all $\mathcal{L}_{at}$ sentences and not some language fragment. I presume that this is what Horsten has in mind, too.

One more thing needs clarification. Horsten says that PKF 'contains' inference rules. This cannot be quite what he means; PKF is made up of *sentences*. I therefore take Horsten to mean that PKF is *closed* under inference rules.[29] In sum, Horsten's third premise becomes

P3″ PKF does not prove universal quantifications into the truth predicate over all $\mathcal{L}_{at}$-sentences, but is closed under inference rules where for any variables $\vec{x}$ within the scope of 'T', $Sent_{at}(\vec{x})$ is a premise, too.

As I explained already above (p. 14), PKF indeed does not contain any sentence of the form

$$\forall x(Sent_{at} \to (\ldots Tx \ldots))$$

Otherwise it could not be sound with respect to Kripke's non-classical model-theory. However, the 'T'-fragment of $\mathcal{L}_{at}$ is closed under inference rules. This is how PKF is defined, but there are derived rules, too. Most prominently, PKF is closed under two inferential analogs of Tarski's T-schema, the rules T-Intro and T-Elim [Halbach and Horsten, 2006, theorem 22].

$$T\text{-Intro}\ \frac{\phi}{T\ulcorner\phi\urcorner} \qquad \frac{T\ulcorner\phi\urcorner}{\phi}\ T\text{-Elim}$$

Thus, there is little about Horsten's third premise that may be questioned; it is simply a fact about the formal theory PKF. By itself, however, this fact also has little philosophical implication. Only together with Horsten's first and second premise PKF allows for his argument for inferentialist deflationism. Especially, only on the assumption that PKF is

[28] It follows Horsten and Halbach's notation [Halbach and Horsten, 2006]

[29] Maybe Horsten had in mind the theory from [Halbach and Horsten, 2006] which consists of *sequents* $\Gamma \Rightarrow \Delta$.

the *best* formal theory we obtain that the deflationist should explain the meaning of the truth predicate by inference rules.

There is a tension in Horsten's position, indeed a fatal one, as I now turn to argue. On one hand, Horsten measures the quality of a formal truth theory by its strength, more precisely, by the range of universal quantifications it proves: the more the better (§1.3). On the other hand, Horsten's argument rests on the best such theory *not* proving quantifications over every sentence (§1.3.4). In other words, Horsten's case for inferentialist deflationism relies on there being an upper bound to the strength of formal truth theory.

Surely, Horsten is careful enough not to commit himself to PKF [Horsten, 2009, p. 22].

> 'PKF (...) only looks good until the better theory comes along. We should surely hold open the possibility that some future stronger inferential truth theory may determine the meaning of the concept of truth even further'

However, Horsten does not want to comment on the specific features of a single theory; he intends to establish a philosophical position. Indeed, he claims that whichever better theory may come, it, too, will be 'inferential', will still not prove general principles.

In consequence, Horsten's argument goes through only if no sound theory proves unrestricted universal quantifications into the truth predicate. This I take to be an overly contentious assumption. In fact, I think it is false.

2 Inferentialist Deflationism Lost

2.1 How To Prove Unrestricted Generalities

In this section I will show that Horsten's argument for inferentialist deflationism is ill-founded. Contrary to his assumption, the best formal truth theory available today does prove 'unrestricted generalities about truth'.

2.1.1 Field's Theory of Truth

Hartry Field has recently elaborated on Kripke's theory [Field, 2003, Field, 2007, Field, 2008]. As in Kripke, the truth predicate is interpreted by a minimal fixed point valuation. Its construction, however, Field iterates in a transfinite revision sequence. This allows him to strengthen the conditional beyond the narrow limits of Kleene logic. For my subsequent discussion it will prove useful to explain Field's work in terms slightly different from his. I clarify the revision-theoretic aspects of his construction in the terminology of §1.3.1.

Field adds to $\mathcal{L}_{at}$ a new binary operator symbol '$\rightsquigarrow$'. Any valuation c of the new sentences $\phi \rightsquigarrow \psi$ can be extended to different valuations v^c of the language $\mathcal{L}_{at\rightsquigarrow}$ as a whole. On the resulting ccpo $\langle \{0, \mathfrak{u}, 1\}^{Sent_{at\rightsquigarrow}}, \leqslant_{\mathcal{V}} \rangle$ an operator K^c can be defined which extends the Kripkean (p. 8) by a single clause (let $Sent_{\rightsquigarrow} = Sent_{at\rightsquigarrow} \backslash Sent_{at}$):

$$K^c(v)(\phi) = c(\phi) \text{ iff } \phi \in \ Sent_{\rightsquigarrow}.$$

Clearly, any such K^c is monotone. Hence, for any valuation $c : Sent_{\rightsquigarrow} \mapsto \{0, \mathfrak{u}, 1\}$ there is a minimal fixed point valuation v^c_{f0}. The new sentences simply keep their values.

To examine the possible interpretations of $\mathcal{L}_{at\rightsquigarrow}$ it therefore suffices to consider the ordering $\mathcal{C} = \langle C, \leqslant_{\mathcal{C}} \rangle$ where $C = \{0, \mathfrak{u}, 1\}^{Sent_{\rightsquigarrow}}$, (the set of valuations c)[30].

On this ordering, an operator $F : C \mapsto C$ can be defined

$$F(c)(\phi \rightsquigarrow \psi) = \begin{cases} 1 & \text{iff } v^c_f(\phi) \leqslant v^c_f(\psi) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This operator F is a *revision rule* in the sense of [Gupta, 1982, p. 38], [Gupta and Belnap, 1993, p. 121]. $F(c)$ is a 'better candidate' [Gupta and Belnap, 1993, p. 121] for an interpretation of $\phi \rightsquigarrow \psi$ than c because now more sentences $\phi \rightsquigarrow \psi$ have designated value (i.e.

[30] The order $\leqslant_c$ is again obtained from the po of the value space $\mathcal{S}$: $c \leqslant_c c'$ iff for every $\phi \in Sent_{\rightsquigarrow}$, $c(\phi) \leqslant c'(\phi)$. $\langle C, \leqslant \rangle$ thus becomes another ccpo [Visser, 2004, lemma 7].

value 1) just in case that the value of the antecedent ϕ is less or equal to the consequent ψ.

Since the value of some formulae may change from 1 to 0 or back , the operator F does not preserve the order on C. Hence, different to Kripke's operator, it does not provide fixed points. Instead, however, it generates *revision sequences* [Gupta and Belnap, 1993, 5C.3]. The general idea is to apply the revision rule again and again, transfinitely many times. Thus, the valuation of the new sentences is continuously improved in the sense just explained.

However, any revision sequence needs some place to start from, a *null valuation* c_0[31]. In Field's case, the new sentences are initially all assigned value $\mathfrak{u}$[32]. Note that trivially, any ϕ has the same value as the sentence that results from ϕ by replacing some subsentence ψ by $T^{\ulcorner}\psi^{\urcorner}$: the null valuation obeys *intersubstitutivity*. This feature is inherited to the minimal fixed point $v_f^{c_0}$ (Field's 'trivial observation' [Field, 2008, p. 243]).

Given the null valuation, F yields a revision sequence, the following transfinite sequence of valuations $(c_0)_\alpha$.

$$(c_0)_0 = c_0 \tag{4}$$

$$(c_0)_{\alpha+1} = F((c_0)_\alpha) \tag{5}$$

$$(c_0)_\lambda = \lim_{\alpha\to\lambda} inf\ (c_0)_\alpha \tag{6}$$

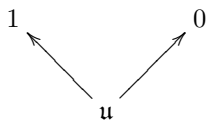
To successor stages the revision rule F is applied. Due to its definition, the values of sentences $\phi \rightsquigarrow \psi$ fluctuate between 1 and 0.

At limit ordinals, then, the valuation is identified with that of the preceding stages that ascribes the most classical values of those that ascribe the least 1s and 0s[33]. In effect, $(c_0)_\lambda$ is identified with that valuation which ascribes 1 (0) to exactly those sentence which from some point below λ on keep this value, but $\mathfrak{u}$ to those sentence whose value continues to fluctuate into the infinite (below λ).

[31] In analogy to Herzberger's term 'null hypothesis' [Herzberger, 1982]

[32] Field allows for, but does not explore, alternative null valuations [Field, 2005, §5][Field, 2008, p. 249]. I will dispense with this.

[33] $\lim_{\alpha\to\lambda} inf\ (c_0)_\alpha = lub\{glb\{(c_0)_\beta | \alpha \leqslant_\mathcal{V} \beta <_\mathcal{V} \lambda\} | \alpha <_\mathcal{V} \lambda\}$ and $\mathcal{V} =$



The intersubstitutivity of c_0 is inherited by all the $(c_0)_\alpha$ (cf. the 'substitutivity' lemma of [Field, 2003, p. 144]).

$(c_0)_\alpha(\phi) = (c_0)_\alpha(\phi(T^\ulcorner\psi^\urcorner/\psi))$ is shown by transfinite induction on α with side-inductions on the complexity of ϕ, similarly to the corresponding lemma about Kripke's theory (p. 9). The base is trivial (see above). Let $\alpha = \beta + 1$ and ϕ be $\chi \rightsquigarrow \xi$ for some atomic χ, ξ. Now either $\psi = \chi$ or $\psi = \xi$, assume $\psi = \chi$.

$$(c_0)_\alpha(\phi) = (c_0)_\alpha(\psi \rightsquigarrow \xi) = F((c_0)_\beta)(\psi \rightsquigarrow \xi)$$

Now since ψ atomic

$$v_f^{(c_0)_\beta}(\psi) = v_f(\psi) = v_f(T^\ulcorner\psi^\urcorner) = v_f^{(c_0)_\beta}(T^\ulcorner\psi^\urcorner) \tag{7}$$

and

$$v_f^{(c_0)_\beta}(\psi) \leqslant v_f^{(c_0)_\beta}(\xi) \text{ iff } v_f^{(c_0)_\beta}(T^\ulcorner\psi^\urcorner) \leqslant v_f^{(c_0)_\beta}(\xi)$$

we have

$$\begin{aligned} F((c_0)_\beta)(\psi \rightsquigarrow \xi) = \begin{Bmatrix} 1 & \text{iff} & 1 \\ 0 & \text{iff} & 0 \end{Bmatrix} &= F((c_0)_\beta)(T^\ulcorner\psi^\urcorner \rightsquigarrow \xi) \\ &= (c_0)_\alpha(T^\ulcorner\psi^\urcorner \rightsquigarrow \xi) = \\ &= (c_0)_\alpha(\phi(T^\ulcorner\psi^\urcorner/\psi)) \end{aligned}$$

For $\psi = \xi$ proceed analogously.

Now let χ and ξ be complex of degree n and assume the claim holds for every $\zeta \in Sent_\rightsquigarrow$ of complexity $\leqslant n$. Again, we can focus on the case that $\phi(T^\ulcorner\psi^\urcorner/\psi)$ is $\chi(T^\ulcorner\psi^\urcorner/\psi) \rightsquigarrow \xi$, the other case is shown in exact analogy. We have

$$v_f^{(c_0)_\beta}(\psi) = v_f^{(c_0)_\beta}(T^\ulcorner\psi^\urcorner) \tag{8}$$

since if on one hand, $\psi \in Sent_{at}$ then (7) holds as above, and if on the other hand $\psi \in Sent_{at\rightsquigarrow}$ then (8) follows from the induction assumption.

$(c_0)_\alpha(\phi) = (c_0)_\alpha(\phi(T^\ulcorner\psi^\urcorner/\psi))$ is now shown as in the induction base.

Finally, assume that α is a limit ordinal. However, since

$$(c_0)_\alpha = \lim_{\beta\to\alpha} inf\ (c_0)_\beta = (c_0)_{\gamma+1}, \qquad \gamma + 1 < \alpha$$

the claim follows by analogous reasoning.

Since no fixed point $(c_0)_\alpha = (c_0)_{\alpha+1}$ can be constructed, some sentences will continue changing their values as the ordinals become bigger and bigger. For many sentences ϕ, however, the value *stabilizes*, i.e. there are ordinals α such that for all ordinals $\gamma \geqslant \alpha$,

$(c_0)_\gamma(\phi) = (c_0)_\alpha(\phi)$. For example, '$T\ulcorner 0 = 1\urcorner \rightsquigarrow T\dot{T}\ulcorner 1 = 1\urcorner$' is ascribed 1 by $(c_0)_1$ and keeps this value thereafter.

Revision sequences such as $(c_0)_\alpha$ eventually enter a *cycle*: there is an *initial ordinal* α_0 such that for every $\beta \geqslant \alpha_0$ it is the case that for any γ there is a $\delta \geqslant \gamma$ such that $(c_0)_\delta = (c_0)_\beta$. In other words, from $(c_0)_{\alpha_0}$ onwards, the valuations *recur* infinitely often [Gupta and Belnap, 1993, 5C.7][34]. Further, it can be proved that if the value of a sentence ever stabilizes, then it has done so at the initial ordinal (ibid., 5C8).

Field now determines the *ultimate value* $c_u(\phi)$ as 1 (0) iff the value of ϕ stabilizes at 1 (0), otherwise $\mathfrak{u}$ [Field, 2003, p. 145][Field, 2008, p. 250f]. c_u occurs in the sequence, in fact, there is an arbitrarily large ordinal Δ such that $(c_0)_\Delta = c_u$ (Field's 'Fundamental Theorem' [Field, 2003, p. 148], [Field, 2008, p. 257ff]).

> The existence of Δ follows from the more general 'Reflection Theorem' of [Gupta and Belnap, 1993, 5C.10] who ascribe it to Herzberger. Apparently, Field has recognized this connection only recently [Field, 2010, fn. 6].
>
> Let δ be a *reflection* ordinal iff δ is $\geqslant$ the initial ordinal α_0 and whenever $\chi \in Sent_{\rightsquigarrow}$ stabilizes at $w \in \{0, \mathfrak{u}, 1\}$ then $(c_0)_\delta = w$. The Reflection Theorem tells now that the class R of reflection ordinals is closed and unbound.
>
> Clearly, for any reflection ordinal δ, $c_u(\chi) = (c_0)_\delta(\chi)$ for those χ that stabilize at 0 or 1. The challenge is to find one such that this holds also for all the χ that do not stabilize. Recall, however, that for any limit ordinal λ, $(c_0)_\lambda(\chi) = \mathfrak{u}$ iff χ does not stabilize below λ.
>
> Now for arbitrary θ, the Reflection Theorem ensures the existence of the least limit ordinal in R above θ. Let Δ be this ordinal. Since Δ is $\geqslant \alpha_0$, any χ unstable in $(c_0)_\Delta$ never reaches a stable value. Therefore, $(c_0)_\Delta(\chi) = c_u(\chi)$ for every χ.

Since every valuation in the sequence obeys intersubstitutivity, $(c_0)_\Delta$ does. Further, it too yields a minimal fixed point and a theory, namely the set of sentences ϕ such that $v_f^{(c_0)_\Delta}(\phi) = 1$. It is this theory that Field eventually endorses; and which I now turn to examine in view of Horsten's argument[35].

[34] Consult also theorem 56 of the helpful [Visser, 2004].

[35] In section 17 of his book Field develops variants of this semantics in an algebraic (§17.1) and a modal setting (§17.2). For the sake of clarity, I confine myself with the (comparably) simpler construction just explained.

2.1.2 Field's Theory Outruns PKF

Since $(v)_f^{c_\Delta}$ is just a Kripke fixed point valuation Field's theory contains Kripke's. Due to the revision theoretic definition of $\phi \rightsquigarrow \psi$, however, it is based on a much stronger logic. For example, Field's theory includes every instance of $\phi \rightsquigarrow \phi$ for as trivial reasons as one might expect[36]. Especially, it contains, for every term t of $\mathcal{L}_{at\rightsquigarrow}$, $Tt \rightsquigarrow Tt$. By intersubstitutivity, Field's theory therefore also contains $Tt \rightsquigarrow T\dot{T}t$. Since the fixed point value of universal quantifications $\forall x\phi(x)$ is the minimum of the values of $\phi(t/x)$, for any term t (K5 from 8), the ultimate value of

$$\forall x(Sent_{at\rightsquigarrow}(x) \rightsquigarrow (Tx \rightsquigarrow T\dot{T}t)) \tag{9}$$

is 1, too.

Horsten rejects the disquotational theory because it does not prove the truth predicate to commute with the connectives (§1.2.3). I emphasized that what he must mean by this is that DT does not contain the universal quantification over every (arithmetical) sentence ((1) on p. 6). Of TC, again, it is an axiom wherefore Horsten takes this to be the better theory of truth. Later, however, he points out that TC does not prove the intuitive self-reflexivity of truth ((2), p. 7). For this reason, Horsten dismisses TC, too, and instead favours PKF that proves (2).

This axiomatization of Kripke's theory, however, does not prove the stronger

$$\forall x(Sent_{at}(x) \rightarrow (Tx \rightarrow T\dot{T}x)) \tag{10}$$

that is the universal quantification over *every* $\mathcal{L}_{at}$-sentence.

Field's theory now proves this principle, which just is a special case of (9). Therefore, if Horsten wants his argument against DT and TC to hold he has to accept that Field's theory outruns PKF.

Now, since Field's theory is the better formal system of truth, Horsten's argument for inferentialist deflationism needs be assessed on its basis. And now it looks much less convincing. As just seen, the theory of the ultimate valuation contains (9). Thus, Field's theory contains universal quantifications into the truth predicate, over every sentence of the extended language. To use Horsten's own phrase, Field's theory proves *unrestricted, general principles* of truth. Thus, Horsten's assumption that the best formal truth theory does not prove such principles is refuted. The argument for inferentialist deflationism breaks down.

[36] Every sentence has always a self-identical value.

In the remainder of this section I will discuss possible objections on Horsten's behalf.

2.2 Discussion

Horsten acknowledges Field's work [Horsten, 2009, p. 24].

> 'The truth theories that are advocated in Soames 1999 and Field 2008 are in the same (partial) spirit as the truth theory that is advocated here. But by insisting that (...) there is a sense in which all the Tarski-biconditonals are correctly assertible (Field), they do not, in my opinion, embrace the Wittgensteinian picture that is defended here as fully as they should.'

Apparently, Horsten takes PKF to be better even than the theory of Field's construction. By the 'Wittgensteinian picture' Horsten presumably refers back to a remark he made earlier in section 5.4 [Horsten, 2009, p. 21]. There, he argues that PKF avoids the sceptical worries attributed to the late Wittgenstein because it does not make general claims about truth. It seems, therefore, as if Horsten rejects Field's theory just because it contains general principles about truth. Clearly, as a response to my criticism this would beg the question.

However, no other argument is found in Horsten's article [Horsten, 2009]. In the meantime, though, Horsten seems to have realized this deficiency since in his forthcoming [Horsten, ta], he discusses Field's work in more detail. There, he rejects it because Field's operator $\rightsquigarrow$ is no natural conditional [Horsten, ta, §10.2.2].

2.2.1 Is $\rightsquigarrow$ a Conditional?

More precisely, Horsten argues that $\phi \rightsquigarrow \psi$ is no adequate formalization of English sentences 'If ... then ...'.

His reason is the following. The schema $\phi \wedge (\phi \rightsquigarrow \psi) \rightsquigarrow \psi$ is not valid in Field's logic[37] [Field, 2008, p. 269]. Any acceptable formalization of the natural language indicative conditional, however, satisfies this object-linguistic schema of *modus ponens*, Horsten assumes.

Unfortunately, this argument does not square well with Horsten's project as a whole. Strong Kleene logic, too, does not validate the modus ponens schema[38] This is no accident. As Field rightly emphasizes [Field, 2008, p. 269], the intersubstitutable truth

[37] Let ϕ be the Curry sentence whose ultimate value is $\mathfrak{u}$, and let ψ be $\overline{0} = \overline{0}$.
[38] Again, let ϕ be $\mathfrak{u}$ and ψ 0.

predicate in combination with this schema give immediate rise to contradiction. Hence, the conditional of Kripke's construction, and therefore that of Horsten's favourite theory PKF, likewise cannot meet what Horsten expects from a conditional.

Moreover, even if Horsten can coherently establish the inadequacy of Field's conditional, I would still not see how this saves Horsten's argument. Recall Horsten's understanding of an 'unrestricted generality about truth' (section 1.4). The conditional he provides is merely an example, all that matters is the universal quantification over every $\mathcal{L}_{at}$-sentence. Assume that $\rightsquigarrow$ is no conditional but some other connective. This does not alter the fact that (9) is an unrestricted generality.

Maybe, however, what Horsten means is rather the following. Field's theory disproves Horsten's assumption only if it is the best theory currently available. Above, I argued that it is because it proves many principles such as (9). This presupposes, though, these principles to capture the compositionality and iteration of the truth predicate of *ordinary discourse*. Maybe it is this assumption that Horsten challenges. Since $\phi \rightsquigarrow \psi$ is no adequate conditional Field's theory does not prove the *real* principles of compositionality and iteration.

Whether Horsten does this move or not, it would not succeed anyway. The reason is simple: Field's theory contains PKF, so it is at least as good as Horsten's preferred theory. Recall that Field merely *extends* the language $\mathcal{L}_{at}$, and that the $\rightsquigarrow$-free fragment of his theory is just a Kripke fixed point theory. Thus, the theory contains also every principle proved by PKF, free of the supposedly dubious new operator but just with the material conditional defined in terms of the strong Kleene operators $\neg$ and $\vee$. Hence, even if (9) and the others do not capture the ordinary discourse principles of truth Field's theory is at least as good as PKF, that is by Horsten's assumption, the best formal theory of truth currently available, and Horsten's reasoning fails.

2.2.2 Deflationism and Model-Theory Revisited

Horsten has a better argument at his disposal. Recall that he dismissed Kripke's theory because it is defined semantically, as the set of sentences that receive designated value in the minimal fixed point model. This model again cannot be defined in the language of the truth predicate but only in a meta-theory. Since the deflationist, however, needs a formal theory that captures ordinary truth talk, no semantical theory can serve the deflationist purpose. For this reason, only the axiomatic theory PKF can be interpreted

deflationistically.

In a similar manner, Horsten could respond to my objection from above[39]. Field's theory, namely, is again defined meta-theoretically. More precisely, he develops his semantics in classical set theory (ZFC) [Field, 2003, p. 166]. Only if axiomatized, it could rival PKF and serve as a counterexample to Horsten's argument.

At this point, Horsten could refer to a result by Philip Welch [Welch, 2008]: Field's theory cannot be axiomatized. Hence, Horsten may argue, it cannot serve to explicate the meaning of the natural language truth predicate. Therefore, it is irrelevant that Field's construction validates unrestricted universal quantifications into 'T'. PKF does not contain any general principle, and since PKF still is the best theory for the deflationist purpose Horsten's argument for inferentialist deflationism is saved.

2.2.3 The Non-Axiomatizability of Field's Theory

Before I reply to this argument, Welch's result requires some further explanation. Strictly speaking, Kripke's theory is not axiomatizable, either (p. 10). It therefore cannot merely be that the set of sentences of ultimate value 1 is not recursively axiomatizable.

Although PKF is not *complete* with respect to Kripke's fixed point models, it still axiomatizes Kripke's theory in the sense that (a) it includes a calculus of strong Kleene logic and (b) it provides rules to construct, given the set of arithmetical truths, the set of $\mathcal{L}_{at}$-sentences of designated value.

Field's theory certainly does not fall behind PKF with respect to (b). As I have shown in §2.1.1, its truth fragment is closed under *intersubstitutivity*. Thus, there is a very simple rule to obtain all the sentences with 'T': simply replace any occurrence of ϕ by $T^{\ulcorner}\phi^{\urcorner}$[40].

However, the theory does not do equally well on (a). Recall that Field's construction differs from Kripke's by its revision theory of $\phi \rightsquigarrow \psi$ (§ 2.1.1). And it was just this that allows his theory to prove general principles such as (9).

Revision theoretic definitions, however, bring with it inevitable complexity. The complexity of revision theoretic definitions, however, inevitably exceeds recursive enumerability. Burgess observed this already with respect to the Gupta Herzberger truth theory

[39] In fact, this is what Horsten showed inclination to in private communication.

[40] In my final section (§3 below) I will specify the inferential character of intersubstitutivity and argue that this suggests an alternative route to inferentialist deflationism.

[Burgess, 1986, § 12]. Welch now showed that Field's construction has a similar effect on the logic of $\rightsquigarrow$.

> The core of his result, in the present terminology, is as follows.
>
> Due to its two case definition (3) (p. 20) Field's operator F yields an arithmetic operator F' where $F'(A)$ is the set of codes of $\rightsquigarrow$-sentences that are value 1 in the fixed point construction based on the valuation c which again is codified by A. Further, Field's null valuation is such that every $\phi \rightsquigarrow \psi$ is ascribed $\mathfrak{u}$. In other words, in the beginning no sentence has value 1. Because the corresponding arithmetic revision sequence thus starts from $\{\varnothing\}$, and due to the *inf lim* limit rule deployed, Field's ultimate valuation amounts to what has become known as an 'arithmetically quasi-inductive' definition [Burgess, 1986, §13].
>
> Now, by a result due to Burgess [Burgess, 1986, 14.1] the set of the codes of sentences that have value 1 in $(c_0)_\Delta$ is Σ_2 definable over the constructible L_ρ, where is ρ the least ordinal α such that for every $\gamma > \rho$, L_γ is Σ_2-reducible to L_α.
>
> Since sets definable over L_{ρ_0} are Δ^1_2 [41] the set of valid sentences $\phi \rightsquigarrow \psi$ is not recursively enumerable. A fortiori, the propositional logic of Field's theory is not axiomatizable, either.
>
> Notice that this negative result would not change if Field had chosen a different null valuation since $\{\varnothing\}$ is already the simplest case [Kühnberger et al., 2005, §4.2].

The propositional logic under which Field's theory is closed is not axiomatizable, different from the strong Kleene logic of Kripke's construction. Consequently, there is no way to construct a proof-theoretic theory that stands to Field's theory as PKF stands to Kripke's.

Therefore, if Horsten is right that only axiomatic theories allow for deflationist interpretation then it seems as if Field's achievements have no relevance for Horsten's case. Then, my criticism would not apply.

In the subsequent sections, however, I will argue that Horsten cannot rule out Field's theory quite as easily.

2.2.4 Does Field's Theory Contain its Own Model-Theory?

Field claims that his theory contains its own model theory [Field, 2003, p. 166]. His model-theoretic construction does not need a stronger meta-theory because it can be conducted within his theory itself.

[41] This fact is cited in [Kühnberger et al., 2005, p. 9] but unfortunately I have not found a proof of it.

To begin with, instead of $\mathcal{L}_a$ of course one may choose a language of set theory $\mathcal{L}_s$ to be extended by 'T' and '$\rightsquigarrow$'. The revision sequence of fixed point models may then be developed on the basis of a model of ZFC set theory [Field, 2007, p. 99].

The resulting theory now contains classical ZFC which provides the means not only to prove the existence of the minimal fixed point models at each stage but also to define the ultimate value of a formula.

Horsten rejects Kripke's theory because it requires a stronger meta-theory and therefore cannot account for the natural language truth predicate. Field's theory, now, contains its own model-theory. It does not need an essentially stronger meta-theory. It seems, therefore, as if Horsten cannot dismiss Field's theory. Instead, he has to acknowledge it as a rival to his own preferred theory PKF. Since Field's theory, however, makes up a counterexample to Horsten's assumption that the best theory does not prove unrestricted generalities, his argument for inferentialist deflationism seems to fail.

There is an obvious worry. How does Field's achievement square with Gödel's second incompleteness theorem: No consistent theory can prove its own consistency? If Field's theory contains its own model theory, then it proves that there are models in which the value of each formula coincides with its ultimate value (the Fundamental Theorem from 23 above.). Thus, it proves that there are models of itself and thereby proves itself consistent.

Field's way out is subtle: The model of ZFC is 'quasi-correct' [Field, 2007, p. 99]. Its domain consists only of the sets of rank up to some inaccessible cardinal[42]. Since the object language now is relativized to this model, especially, the quantifiers of $\mathcal{L}_{st\rightsquigarrow}$ are restricted to the set of sets below the inaccessible ordinal. Thus, they do not range over the *real* set theoretic universe, the cumulative hierarchy.[43] Consequently, however, the model Δ, too, cannot be the intended model of Field's theory of truth.

Field does not ignore this fact but endorses it as a feature of his view. Preservation of value 1 cannot be *real* validity [Field, 2008, p. 277]. Consequently, if Field's relativized theory proves the Fundamental Theorem then it proves merely its *model-theoretic* consistency. Field acknowledges that this does not ensure real consistency [Field, 2008, p. 67] but accepts this limitation. The value of the model-theory consists rather in that it (ibid.)

[42] That such models validate ZFC is a textbook result, see e.g. [Jech, 2002, 12.13].

[43] A similar relativization of the set theoretical object language Field considers already in his discussion of Kripke's theory [Field, 2008, p. 63]. There he calls it a 'misinterpretation'.

> '(...) provides an easy way to figure out which inferences involving truth are legitimate in the truth theory provided by the construction.' [44]

Field confines himself with this.

But one need not be that modest. Indeed, the relevance of a model theory such as Field's may well be questioned. Traditionally, model theory is seen as proving *real* consistency because it determines the meaning of the linguistic items under investigation, especially of the connectives and quantifiers. Field willingly abandons this justificatory role of semantics [45].

The objection on Horsten's behalf from the preceding section therefore holds at least so far: Field's theory does not contain a model theory in the received sense of the term. But, the role of model theory may eventually be a methodological question that in the long run will be settled by practise. Horsten would be ill-advised, I think, if he did not base his case on a firmer grounding.

As I will argue in the subsequent section, however, the objection from model theory may be developed in a different manner that also seems in line with Horsten's reasoning elsewhere in his article.

2.2.5 Revenge

In the vicinity of his remarks about the Wittgensteinian character of PKF [Horsten, 2009, p. 21], Horsten reminds the reader of a problem Kripke himself discussed with respect to his construction [Kripke, 1975, p. 714].

The paradoxical sentences do not receive a designated value in the fixed point model This fact, however, cannot be expressed in the object theory on pain of *revenge*. Assume it could. By diagonalization there was a sentence provably equivalent to itself not having designated value. Since the fixed point models are constructed in classical set theory, any sentence either has designated value or not, in particular this sentence. And we're back in paradox.

Hence, there is a semantic fact about the theory that it cannot express itself. It is only by the means of the meta-theory that the status of the paradoxical sentences can be expressed. For this reason, Kripke eventually admits the (ibid.)

[44] This remark refers to Kripke's theory [Field, 2008, §3.2]. To its discussion, however, he later refers back in order to justify his choice of a quasi-correct model [Field, 2008, p. 257].

[45] Here I summarize worries raised in [Priest, 2007b, Priest, 2007a, Priest, 2010]

(...) necessity to ascend to a metalanguage (...).

A theory really avoids the need of a meta-theory only if it can express its own semantics, only if it is *semantically self-sufficient*.

A meta-theory can determine the semantic values of every sentence because it can talk about the theory's intended models. This, now, is precisely what the model theory included in Field's theory is not supposed to do. The response on Horsten's behalf thus can be reformulated without a commitment to the justificatory role of model theory. Field's theory does not disprove Horsten's argument for inferentialist deflationism because its semantics, on pain of revenge, is fully determined only in a theory of its intended models. Such a meta-theory, however, it cannot include.

Nonetheless, Field claims his theory to be revenge-immune. What does he mean, if it cannot be that his theory constructs its own intended models? Kripke's revenge problem may be formulated differently. If the minimal fixed point model is the theory's intended model, then the fact that the Liar sentence λ does not have value 1 in it amounts to λ not being *determinately* true.

From the object theory's point of view revenge occurs if one tries to say that the paradoxical sentence κ is not *determinately* true. In Kripke's theory, as well as in PKF, there is no consistent way to obtain such an operator. Field claims to have done better. The conditional he has defined allows for an operator D of *determinate truth* [Field, 2003, p. 157].

$$D\phi :\leftrightarrow (\top \rightsquigarrow \phi) \wedge \phi$$

For a paradoxical sentence such as the liar λ, the ultimate value of $D\lambda$ is 0.[46] Hence, $\neg D\lambda$ becomes 1, and Field's theory expresses the status of the liar sentence: It's not determinately true.

Why does the corresponding Liar λ' where $\lambda' \leftrightarrow \neg D\lambda'$ not lead back to paradox? It is not in the theory: $v_f^{(c_0)_\Delta}(D\lambda')$ stabilizes at $\mathfrak{u}$ [Field, 2003, p. 159]. Thus, the ultimate value of λ' is $\mathfrak{u}$, as well.

However, the determinately operator D can be iterated; and for a similar reasoning as above, $DD\lambda'$ is in the theory. Therefore, the theory is capable also to express the status of this strengthened liar: λ' is not *determinately determinately true*.

[46] At any fixed point model α, $v_f^{(c_0)_\Delta}(\lambda) = \mathfrak{u}$. Hence, $v_f^{(c_0)_\Delta}(\top \rightsquigarrow \lambda)$ stabilizes at value 0. Since conjunction follows the strong Kleene rules, $(\top \rightsquigarrow \lambda) \wedge \lambda$ gets an ultimate value of 0, too.

The iteration of D can be continued into the transfinite: Field shows that no strengthened Liar $\neg \underbrace{DD\ldots D}_{\alpha\times}\lambda^{\alpha}$ is in the theory but that for each of these, the theory contains $\neg D\underbrace{DD\ldots D}_{\alpha\times}\lambda^{\alpha}$ [Field, 2003],[Field, 2008, p. 237].

On this basis, Field claims to have provided a 'revenge-immune' truth theory. And indeed, Field's theory now seems capable of expressing the semantic value of every possible sentence, In the benign cases, it suffices to use the truth predicate 'T'; the values of paradoxical sentences again, that lie somewhere between 1 and 0, can be described by means of the corresponding iteration of 'D'.

Revenge problems as with Kripke's theory reveal expressive limitation and thus show that a meta-theory is needed to express the whole semantics of the theory. By having defined a workable determinacy operator, Field has made a strong case that his theory avoids revenge and thus is not haunted by Tarski's ghost. Consequently, Field's theory makes up a valid counterexample to Horsten's argument for inferentialist deflationism, which therefore need be rejected.

However, I admit that Field's theory is discussed controversially[47]. Especially its alleged revenge-immunity has been questioned. Just to mention one prominent opponent, Graham Priest argues that Field's theory, although it can express every single level of determinacy, fails to express the general notion of determinacy [Priest, 2010, pp. 123f]. The operator D cannot do this job, since for every iteration $\underbrace{DD\ldots D}_{\alpha\times}$ there is a determinately true sentence that does not fall under the notion expressed:

$$\neg D\underbrace{DD\ldots D}_{\alpha\times}\lambda^{\alpha}$$

Going to limit ordinals does not help either since Field's hierarchy is and needs be extended into the transfinite.

Moreover, Priest argues, the failure is a matter of principle. Assume some operator G did apply to every determinately true sentence. The corresponding diagonalization $\neg G\gamma$ then must not be determinately true on pain of contradiction. But since this is determinately so, and G should apply to *every* determinately true sentence, $G\gamma$ is in the theory, too. Contradiction. In sum, Priest argues that Field after all does face a revenge objection. There is no way for his theory to talk about every determinately true sentence.

[47] A fine selection of contributions is found in the 147th volume of the *Philosophical Studies*.

Field makes some effort to check Priest's challenge [Field, 2007, part 4] [Field, 2008, pp. 343nn.]. He argues that the contradiction to which the operator G gives rise is not a defect of his theory. Instead, general determinacy itself is an incoherent notion. Thus, however, the debate has reached stalemate. Therefore, I prefer not commit myself to Field's theory being in fact revenge-immune. Fortunately, I need not. My objection does not depend on the expressive power of Field's theory, as I shall argue in the next section.

2.3 The Supposed Semantic Self-Sufficiency of Ordinary Discourse

Horsten's case for inferentialist deflationism depends on specific features of Kripke's theory, respectively its axiomatization PKF and, crucially, on his assumption that any theory better than PKF shares these properties.

I argued that Field's work provides a counterexample to Horsten's claim that the best formal theory does not contain universal quantifications into the truth predicate. I considered a response on Horsten's behalf: Field's achievements are irrelevant for the deflationist because it can be determined only by model-theoretic means, and therefore cannot be applied to ordinary discourse.

Since Field's theory does contain its own model theory, although it cannot define its intended models, I rephrased Horsten's argument against model-theory as a *revenge* objection. I discussed Field's claim that his system is revenge-immune, but could not find his argument conclusive. Revenge may indeed be a problem for Field.

But why should it be a problem for the deflationist? The alleged expressive limitations of Field's theory would disallow its deflationist interpretation only if it blocked its application to ordinary discourse. This again presupposes ordinary discourse not to suffer from similar restrictions.

This idea is not uncommon. It may be traced back to Tarski himself, who held that 'if we can speak meaningfully about anything at all, we can speak about it in colloquial language' [Tarski, 1956, p. 164]. Tarski did not yet distinguish sharply between languages and theories. Horsten, too, talks of the (lack of a) meta-*language* of *English*. Nonetheless, all theories considered in the course of the present discussion are formulated in the *same* language, a first order formal language of arithmetic, extended by a truth predicate 'T'. Field adds a connective, but what matters is the resulting theory. Therefore, I take the issue to be about theories, and therefore continue talking about *ordinary discourse*. Tarski's universality thesis thus becomes the claim that ordinary discourse is the strongest

theory. This would indeed bolster Horsten's implicit assumption.

However, the *universality* of ordinary discourse, taken at face value, is plainly false. The theory PKF in the language $\mathcal{L}_{at}$ is not part of ordinary discourse in English, French or Chinese [Gupta and Belnap, 1993, p. 257]. Fairer, maybe, it is to take Tarski to claim an indefinite extensibility of natural languages: 'anything whatsoever can be expressed in them once suitable resources are added' (ibid.). This again is certainly right but holds just as much of any language: especially of all the formal theories for which we have set up well functioning meta-theories. This weaker reading therefore cannot justify Horsten's rejection of semantical theories, either.

The most promising response to my objection from §2.1.2 is a *revenge* objection. Field's theory, it goes, cannot develop all of its own semantics. In order to rule out my counterexample Horsten therefore must presume that ordinary discourse is *semantically self-sufficient*.

However, this assumption is overly contentious, as I will argue in the next section[48]

2.3.1 Semantic Self-Sufficiency

Semantic self-sufficiency presupposes that the semantics of a natural language can be formulated in this very language. Different to the formal languages considered above, however, a natural language is an essentially *indeterminate* object. Natural languages gain and lose vocabulary and also their grammar changes continuously. Is ordinary discourse supposed to provide the semantics of every such stage? This is absurd. It is impossible to determine even the syntax of future stages of English. Also, its features become less and less known the further one looks into the past.

The only reasonable approach, therefore, is to consider the current stage of its language. This again commits Horsten to the semantic self-sufficiency of our discourse *today*. If only speaking the English of the year 2099 we could interpret it fully, what would be the difference to meta-theoretical reasoning?[49] This could not justify Horsten's rejection of semantical truth theory[50].

[48] I adopt the forceful reasoning found in [Gupta and Belnap, 1993, pp. 257n], and especially [Gupta, 1997, pp. 439n].

[49] Some authors do understand semantic self-sufficiency in this weak sense, e.g. [Simmons, , pp. 13n]. Maybe they do not consider that as the common reasoning of 2099, a meta-theory may be nothing more than an extension, or development of the object-theory.

[50] In his forthcoming book, Horsten seems to agree with me on this [Horsten, ta, §2.3].

Now, however, we need to ask: what justifies the assumption that today we have a complete semantics of our own reasoning? Horsten does not hint at why he thinks so but elsewhere, an interesting argument is found. Vann McGee justifies (the assumption of) semantic self-sufficiency from a broadly naturalist stance [McGee, 1994, p. 628]. Whoever subscribes to the view that human life is '(...) amenable to scientific understanding (...)' [McGee, 1994, p. 628] must especially hold that the semantics of our common reasoning is comprehensible to us.

In order to reject meta-theoretical truth theory, however, this line of thought presupposes that ordinary speakers *now* have this understanding. Now, Gupta distinguishes between two ways this may be meant [Gupta, 1997, pp. 441n].

> In one sense, it means simply the ability to understand and use the language. In this sense it is tautological that English is comprehensible by English speakers. And nothing much follows from this triviality. In the other sense, 'comprehensibility' means the ability to give a systematic theory of English.

Therefore, the thesis of semantic self-sufficiency can hold only if ordinary speakers are capable, today, of providing a complete, scientific semantics of their reasoning. This, however, seems plainly false[51]. I can only agree with Gupta when he concludes [Gupta, 1997, p. 422].

> The philosophical underpinnings of semantic self-sufficiency need to be carefully considered before it is used as a criterion of adequacy on theories of truth.

Moreover, it is not wise anyway for Horsten to commit himself to semantic self-sufficiency as a necessary requirement on formal truth theory. PKF, namely, is not semantically self-sufficient, either.

Being a sub-theory of Kripke's, it is likewise not able to express that the liar sentence has value $\mathfrak{u}$. In fact, since it is a subtheory of Field's (section 2.2.1) it is at least as expressively limited. Horsten claims that PKF avoids revenge because it [Horsten, 2009, p. 21]

[51] Matti Eklund recently has provided a helpful overview on the positions in this question, in which he agrees with me that '(...) the arguments for semantic self-sufficiency are unpersuasive (...)'[Eklund, 2007, p. 59]

makes no claim concerning the truth value of the liar sentence.

What solution, however, is made up by such quietism? Either, he means that revenge is a problem only for meta-theoretically determined theories. Then, however, Horsten would beg the question against Field. Or, he frankly admits that PKF is likewise not semantically self-sufficient. In this case, however, Horsten could not reject Field's theory because of its expressive limits, on pain of losing the basis of his own argument for inferentialist deflationism.

In the end, therefore, I do not see a way for Horsten to respond to my objection from §2.1.2. His argument for inferentialist deflationism fails in view of Field's recent achievements.

3 Inferentialist Deflationism Regained?

Horsten's argument from formal truth theory has failed. Does this make inferential deflationism untenable? I do not think so. In the remainder of the essay I would like to suggest an alternative argument for it.

3.1 Intersubstitutivity

As I explained back in section 1, Horsten argues for PKF on the basis of three intuitions. First, in ordinary discourse we find sentences 'ϕ is true if and only if ϕ' trivially true. This *disquotational* intuition motivated the theory DT of §1.2.3. Second, if I claim that ϕ is true and ψ is true, too, then I am also committed to accept that 'ϕ and ψ' is true, and similarly for the other connectives and quantifiers. The degree of this commitment I have called the *compositional* intuition (p. 1.2.3).

Finally, Horsten emphasizes the importance of a third intuition (p. 7). If you say that ϕ is true, you can be held to have claimed, too, that 'ϕ is true' is true. This may seem just a special case of disquotation. In fact, however, neither DT nor TC can account for the *iterative* intuition.

Horsten now argued that only inferential theories such as PKF could accommodate all three intuitions. In the preceding section, I argued against this assumption. Now, I would like to look at it from a different angle.

The theory of the minimal fixed point accommodates all three intuitions (within the narrow limits of strong Kleene logic) because it obeys the principle of *intersubstitutivity* (p. 9), and the same holds for the theory KFS of §1.3.2. Horsten's favourite theory PKF can therefore be seen as the attempt to approximate, by proof-theoretic means, this very principle. Furthermore, that Field's theory again outruns PKF can also be traced to the fact that it allows for full intersubstitutivity. It reduces (9) from p. 24 to the triviality of a logical truth. I therefore think that my discussion from the preceding sections indicates that it is this principle which enables formal theories to capture ordinary truth talk.

However, the discussion of the preceding section has likewise shown that any philosophical interpretation of formal truth theory must be treated with caution. Fortunately the strength of the intersubstitutivity principle becomes clear also in an informal context. Once we assume that any sentence ϕ is always interchangeable with 'ϕ is true', commitment to ϕ becomes commitment to 'ϕ is true' and vice versa (disquotation), the acceptance of 'ϕ is true and ψ is true' is acceptance of "ϕ and ψ' is true' (compositionality)

and saying 'ϕ is true' is just to say "ϕ is true' is true' (iteration).

Horsten assessed formal theories on the basis of three different intuitions. I would now like to suggest that they have a common basis. Intersubstitutivity is the fundamental principle that underlies our reasoning with truth.

If this analysis is correct then the deflationist may achieve an adequate account of ordinary truth talk if she just assumes the intersubstitutivity of the truth predicate. Furthermore, the immediate connection between intersubstitutivity and disquotation, compositionality and iteration holds independent of model- or proof-theoretic details of one's favourite truth theory, in fact does not rely on any formal account. Therefore, the deflationist need not commit herself to contentious claims such as Horsten assumes in his reasoning.

I do not claim this to be a very innovative proposal. The idea seems to have been in the wind for some time. Field's work and the important role he ascribes to intersubstitutivity may well be motivated by his deflationist views [Field, 2001]. JC Beall recently sketched a very similar account of deflationist truth [Beall, 2010][52]. However, even if deflationists tend to confine themselves with general lines the idea needs further specification.

But how do we formulate the intersubstitutivity of truth? The T-schema, the compositional axioms of TC, the axioms of KF and the rules of PKF may all be seen as approximations to it, but do not capture the idea. Given an operator such as Field's $\rightsquigarrow$, intersubstitutivity can be formulated as an axiom [53]

$$\forall x \forall y (Tx \leftrightsquigarrow T\dot{Sb}(\dot{T}(y), y, x))$$

However, Kripke's theory shows that one need not elaborate on the propositional logic and buy into its non-axiomatizability to obtain intersubstitutivity. Already in strong Kleene logic intersubstitutivity can be formulated as two rules of *substitution*:

$$Sb\mathrm{I}\ \frac{\psi}{\psi(T^{\ulcorner}\phi^{\urcorner}/\phi)} \qquad \frac{\psi(T^{\ulcorner}\phi^{\urcorner}/\phi)}{\psi}\ Sb\mathrm{E}$$

[52] Beall builds a *paraconsistent* theory of truth on the principle of intersubstitutivity, an approach Horsten does not consider at all in his discussion of formal truth theories. I have followed him in this.

[53] Where $\dot{Sb}(^{\ulcorner}\phi^{\urcorner}, ^{\ulcorner}\psi^{\urcorner}, ^{\ulcorner}\chi^{\urcorner})$ encodes the syntactical operation $\chi(\phi/\psi)$. A more precise notation would make it a four place operator to explicate *which* occurrences are replaced. Since all these are just equivalent, though, this would be an unnecessary complication.

In an important sense soon to be specified, these rules constitute generalized introduction and elimination rules for 'T': not merely ϕ alone but any occurrence of ϕ allows for $T^{\ulcorner}\phi^{\urcorner}$ (SbI) and any occurrence of $T^{\ulcorner}\phi^{\urcorner}$ again allows for ϕ (SbE).

Since intersubstitutivity does not need to be formulated as an axiom, I would like to suggest that the inferential formulation is the more basic. Suddenly, *inferentialist* deflationism again seems within reach.

Clearly, however, my reasoning so far has again been rather interpretative than deductive. It can motivate but certainly not establish inferentialist deflationism. More is needed. In the subsequent, final section I will therefore sketch a direct argument that the substitution rules exhaust the meaning of the truth predicate.

3.2 A Direct Way to Inferentialism about Truth

3.2.1 Meaning-Constitutive Rules

When a set of rules determine the meaning of an expression, has been investigated by proof-theorists for decades. I connect with one line of thought that has gained momentum recently [Read, 2000, Francez and Dyckhoff, 2009, Read, 2010]. It can be traced back, however, to the founder of proof-theory, Gerhard Gentzen. He wrote [Gentzen, 1969, p. 80]

> The introductions represent, as it were, the 'definitions' of the symbols concerned, and the eliminations are no more, in the final analysis, than the consequence of these definitions.

After the war, it was Paul Lorenzen who returned to the question how a set of rules may fully determine the use of a given expression $\star$ [Lorenzen, 1955, p. 30]. More precisely, he discussed how to ensure that any sentence with $\star$ is only obtained by the given rules. Lorenzen's answer is much in line with Gentzen's idea: Whatever follows from any conclusion of the *introduction* rules must follow also directly from their premises. Especially, proper *elimination* rules are mere consequences of the introductions, just as Gentzen suggested[54]. Since now, given the introduction $\phi \Rightarrow \psi$, any elimination rule

[54] Lorenzen formulated the principle for sequent-calculi. Only Dag Prawitz applied Lorenzen's principle to the calculus of natural deduction [Prawitz, 1965, p. 33]. Consult also the helpful [Moriconi and Tesconi, 2008].

$\psi \Rightarrow \chi$ requires $\phi \Rightarrow \chi$, and especially $\phi \Rightarrow \psi$, Lorenzen called this rule-generating principle the 'inversion principle'.

*Sb*I and *Sb*E clearly obey the inversion principle[55]. The consequence of an elimination of $\psi(T^{\ulcorner}\phi^{\urcorner})$ follows trivially from the premises of its introduction; it is just this very ψ. Therefore, if the proof-theoretic tradition starting from Gentzen is right that the inversion principle ensures a set of rules to fully determine a symbol's meaning then the meaning of 'T' is exhausted by inference rules.

It may seem that such an argument establishes inferentialism about truth but does not suffice for a deflationist account. In this case, inferential *deflationism* would still not have been achieved. In fact, however, once we have established inferentialism about truth the step to deflationism is made easily.

If the meaning of an expression is captured by inference rules then it need not be interpreted as an item or subset of the domain of discourse. It can be meaningfully used independently of what there is. Ontological neutrality simply follows from the inferentialist '(...) idea of privileging *inference* over *reference* in the order of semantic explanation (...)' [Brandom, 2000, p. 1]. Inferentialism about truth thus implies that truth talk is ontologically neutral. Truth becomes a *light* notion just in the deflationist's spirit.

3.2.2 Discussion

Let me briefly consider two possible objections against this proposal. First, it may be argued that the substitution rules do not exhaust the meaning of 'T' because they presuppose, for any sentence ϕ, the existence of the name $^{\ulcorner}\phi^{\urcorner}$. In the end, '$T$' is not an operator but a first order *predicate*. As is well known, however, the quotation device $^{\ulcorner}\cdot^{\urcorner}$ is only defined in a theory of syntax at least as strong as primitive recusirve arithmetic, (see fn 3 above). In its absence, the rules *Sb*I and *Sb*E do not even accommodate the simple disquotational intuition because we may prove ϕ without being able to obtain $T^{\ulcorner}\phi^{\urcorner}$. Hence, the objection goes, the intersubstitutivity rules alone do not make 'T' a truth predicate. wherefore no inferentialism about truth can be based upon them.

Although it rightly reminds of an often neglected fact, I think that this objection misfires. For one, any of the attempts to capture the notion of truth I have considered in the

[55] See [Read, 2000, p. 127] for a similar argument, but with respect to simpler rules. Read ascribes the idea to [Prawitz, 1994, p. 347].

course of the present paper involves the same implicit assumption. Especially, Horsten's favoured theory PKF is assumed to include full PA (§1.3.3). Now, his argument for inferentialist deflationism has failed, but not because of the number theoretical commitments of his theory.

The need for syntax does not contradict with my thesis that the rules of intersubstitutivity exhaust the meaning of the truth predicate. In chess, the rules for the knight tell you every move it can make; and whenever, in a match, the piece has been moved, it has been moved according to these rules (otherwise you would not play chess). In this respect, the rules fully determine the *meaning* of the piece. Clearly, however, the rules can be applied only in the context of a game. Thus, the usage of the knight presupposes the rules for all the other pieces[56]. I suggest to think analogously of the rules $Sb\mathrm{I}$ and $Sb\mathrm{E}$. They determine every move you can make with the truth predicate, but for a full theory of truth you also need to know how to talk about your language.

However, I admit that this perspective on rules and their application is not uncontroversial. Fortunately, I have available an alternative response. Since Robinson arithmetic Q provides the required resources as well as is finitely axiomatizable, the implicit assumption of syntax can be made explicit.

$$Sb\mathrm{I}' \frac{Q \wedge \psi}{Q \wedge \psi(T^{\ulcorner}\phi^{\urcorner}/\phi)} \qquad \frac{Q \wedge \psi(T^{\ulcorner}\phi^{\urcorner}/\phi)}{Q \wedge \psi} Sb\mathrm{E}'$$

These rules, although necessarily more cumbersome, still clearly obey the inversion principle, and therefore make up a complete inferentialist account of truth.

The second objection I would like to consider is based on a view widely held among logical inferentialists, namely that a set of rules can only be considered meaning constitutive if they their addition to a given theory remains *conservative*[Dummett, 1991, pp. 217, 250]. This means, the extended theory must not prove any statements in the 'T'-free language that the original theory did not already prove, too.

On this basis it may be argued that intersubstitutivity is not conservative over the arithmetic base theory. If one adds $Sb\mathrm{E}$ and $Sb\mathrm{I}$ to PA then one is able to prove the

[56] And the list of presuppositions can be continued: you need an opponent player (human or electronic), a chess board and pieces, and so on.

global reflection principle for PA[57]

$$\forall x(Prov_{PA}(x) \rightarrow Tx)$$

which together with *Sb*I allows to prove $\neg Prov_{PA}(\ulcorner 0 = 1 \urcorner)$, that is the consistency of PA. For this reason, the objection goes, the substitution rules cannot be meaning-constitutive for the truth predicate.

I do not think that this objection from non-conservativeness defeats my proposal of an inferentialist deflationism on the basis of intersubstitutivity. On one hand, there are good reasons simply to accept the non-conservativeness of the rules. Any theory that accommodates the compositional intuition, that is, any theory at least as strong as TC proves the global reflection principle. In fact, before turning to inferentialist deflationism, Horsten argues at length for the non-conservativeness of truth [Horsten, 2009, §§ 3-4].

Moreover, the conservativeness constraint initially applied only to logical rules. If we aim for an inferentialist account of truth, this constraint may well not carry over. Instead, it is the *inversion principle* which ensures that the substitution rules exhaust the truth predicate's meaning, independent of their non-conservativeness[58].

On the other hand, I do not need to commit myself to inferentialism beyond conservativeness. The consistency proof of the base theory presupposes that the object-linguistic induction schema is extended to formulae which contain the new predicate 'T'. Some deflationists, prominently Hartry Field [Field, 1999], reject this move. In this case, the intersubstitutivity rules indeed are conservative, for example, the theory KFS is conservative over PA [Field, 2008, p. 66].[59]

Nonetheless, a conclusive case certainly goes beyond the scope of the present study. I merely wish to suggest that although Horsten's argument for inferentialist deflationism has failed, the idea of combining these two strands of contemporary philosophy is no lost cause. However, it needs to be established differently from how Horsten does it: not

[57] For any axiom ϕ we prove $T\ulcorner\phi\urcorner$, and since we can prove that *modus ponens* and universal generalization preserve truth, we can infer by induction that every provable sentence is true. This induction on the length of a proof is easily arithmetized and thus conducted within our truth theory itself. Notice that this reasoning requires the rules of *intersubstitutivity* and does not go through with the weaker T-Intro and T-Elim (p. 18).

[58] That the inversion principle may be defended even in view of non-conservative rules has been argued in [Read, 2010]. In his terminology, rules may be 'general-elimination harmonious' even if not conservative.

[59] This extends to Field's strengthened $\rightsquigarrow$-logic: [Field, 2008, p. 263].

from contentious assumptions about truth theory but by a careful examination of the supposedly meaning-constitutive inference rules.

Conclusion

Pairing deflationism about truth with an inferentialist account of the truth predicate provides an attractive opportunity to specify and strengthen the deflationist position. Horsten has derived *inferentialist deflationism* from an interpretation of formal truth theory. The present paper showed that this approach does not succeed, and sketched an alternative.

To lay my critique on a firm grounding I first analyzed in detail Horsten's assumptions and their logical background (§1). My examination showed that his argument presupposes the prospects of formal truth theory to be limited: the theory that proves the most universal quantifications into the truth predicate does not prove *unrestricted* quantifications. To this claim I advanced a counterexample (§2.1.2). Field's theory proves more principles of truth then Horsten's favourite PKF. But it also proves unrestricted universal quantifications, disproving Horsten's assumption (§2.2.1).

I then turned to discuss possible responses to my objection. First (§2.2.1), I considered a worry about Field's proposal that Horsten raises elsewhere [Horsten, ta]. He argues that sentences $\phi \rightsquigarrow \psi$ cannot be regarded as adequate formalizations of natural language conditionals. This reasoning, however, cannot rule out Field's theory as a counterexample to Horsten's assumption, for two reasons. First, the criticism equally applies to the conditionals of PKF, second, Horsten's classification of truth theories does not require them to prove conditional principles.

Consequently, I focused on a different response which I found motivated by his treatment of Kripke's fixed point model theory. In §2.2.2 I argued on Horsten's behalf that Field's work is irrelevant for the deflationist because the theory is not axiomatizable.

However, since at least in some sense of the term Field's theory contains its own model theory (§2.2.4) this argument needed specification. Similar to the *revenge* objections that pervade the discussion of non-classical solutions to the paradoxes, the response is better phrased as accusing Field of expressive limitation (§2.3).

Fortunately, I did not have to settle this controversial matter because Horsten's rejection of semantical theories is flawed on different grounds. It presupposes the *semantic self-sufficiency* of ordinary discourse. I explained why this assumption is a contentious empirical claim as well as at odds with Horsten's preference for PKF (§2.3.1). I concluded that Horsten's argument for inferentialist deflationism fails.

In the final section, I then turned to explore an alternative approach. I traced back

the truth theoretic strength both of PKF and of Field's theory to the principle of *intersubstitutivity*, and argued that it also underlies ordinary reasoning with truth. Moreover, intersubstitutivity is best captured by two inference rules (p. 38). I took this as strong evidence for inferentialist deflationism, but emphasized that it does not yet provide sufficient justification.

As an example of how such justification may be found I finally took up a theme from contemporary proof-theoretic semantics [Read, 2000, Francez and Dyckhoff, 2009, Read, 2009, Read, 2010]. Inferentialists about the logical constants have long investigated under which conditions a set of rules exhaust the meaning of a connective or quantifier. The most promising of these attempts, Lorenzen's *inversion principle*, is also easily applied to the rules of the intersubstitutivity of truth (§3.2.1), even in view of their non-conservativeness and their reliance on a background theory of syntax (§3.2.2). Thus, I did not only sketch an argument for inferentialist deflationism independent of Horsten's controversial interpretation of truth theories, but have also built a bridge between this project and the front line of inferentialist research.

References

[Beall, 2010] Beall, J. (2010). *Spandrels of Truth*. Oxford University Press, New York.

[Beall and Armour-Garb, 2005] Beall, J. and Armour-Garb, B. (2005). *Deflationism and Paradox*. Clarendon Press, Oxford.

[Brandom, 2000] Brandom, R. B. (2000). *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, Cambridge, Massachusetts.

[Burgess, 1986] Burgess, J. P. (1986). The Truth is Never Simple. *The Journal of Symbolic Logic*, 51(3):663 – 681.

[Cantini, 1989] Cantini, A. (1989). Notes on Formal Theories of Truth. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 35(2):97 – 130.

[Dummett, 1991] Dummett, M. (1991). *Proof-Theoretic Justifications of Logical Laws*, chapter 11. Duckworth, London.

[Eklund, 2007] Eklund, M. (2007). The Liar Paradox, Expressibility, Possible Languages. In Beall, J., editor, *Revenge of the Liar*, pages 53 – 77. Oxford University Press, Oxford.

[Feferman, 1991] Feferman, S. (1991). Reflecting on Incompleteness. *The Journal of Symbolic Logic*, 56.

[Field, 1999] Field, H. (1999). Deflating the Conservativeness Argument. *The Journal of Philosophy*, 96(10):533–540.

[Field, 2001] Field, H. (2001). *Deflationist Views of Meaning and Content*. Oxford University Press, Oxford.

[Field, 2003] Field, H. (2003). A Revenge-Immune Solution to the Semantic Paradoxes. *Journal of Philosophical Logic*, 32:139 – 17.

[Field, 2005] Field, H. (2005). Variations on a Theme by yablo. In Beall, J. and Armour-Garb, B., editors, *Deflationism and Paradox*. Oxford Universit Press, Oxford.

[Field, 2006] Field, H. (2006). Truth and the Unprovability of Consistency. *Mind*, 115.

[Field, 2007] Field, H. (2007). Solving the Paradoxes, Escaping Revenge. In Beall, J., editor, *Revenge of the Liar: New Essays on the Paradox*, pages 78 – 144. Oxford University Press, Oxford.

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

[Field, 2010] Field, H. (2010). Precis of *Saving Truth from Paradox*. *Philosophical Studies*, 147:415–420.

[Fitting, 1986] Fitting, M. (1986). Notes on the Mathematical Aspects of Kripke's Theory of Truth. *Notre Dame Journal of Formal Logic*, 27(1).

[Francez and Dyckhoff, 2009] Francez, N. and Dyckhoff, R. (2009). A Note on Harmony.

[Gentzen, 1969] Gentzen, G. (1969). Investigations concerning Logical Deduction. In Szabo, M., editor, *The Collected Papers of Gerhard Gentzen*, pages 68 – 131. North-Holland, Amsterdam.

[Gupta, 1982] Gupta, A. (1982). Truth and Paradox. *Journal of Philosophical Logic*, 11(1):1–60.

[Gupta, 1997] Gupta, A. (1997). Definition and Revision: A Reponse to McGee and Martin. *Philosophical Issues*, 8: *Truth*:419–443.

[Gupta and Belnap, 1993] Gupta, A. and Belnap, N. (1993). *The Revision Theory of Truth*. MIT Press, Cambridge, MA.

[Halbach, 1996] Halbach, V. (1996). *Axiomatische Wahrheitstheorien*. Akademie Verlag.

[Halbach, 2009] Halbach, V. (2009). Reducing Compositional to Disquotational Truth. *The Review of Symbolic Logic*, 2(4).

[Halbach and Horsten, 2003] Halbach, V. and Horsten, L. (2003). Contemporary Methods for Investigating the Concept of Truth: An Introduction. In *Principles of Truth*. ontos, Frankfurt, London.

[Halbach and Horsten, 2005] Halbach, V. and Horsten, L. (2005). The Deflationist's Axioms for Truth. In Beall, J. and Armour-Garb, B., editors, *Deflationism and Paradox*, pages pp. 203–217. Clarendon, Oxford.

[Halbach and Horsten, 2006] Halbach, V. and Horsten, L. (2006). Axiomatizing Kripke's Theory of Truth. *The Journal of Symbolic Logic*, 71(2).

[Herzberger, 1982] Herzberger, H. (1982). Notes on Naive Semantics. *Journal of Philosophical Logic*, 11:61 – 102.

[Horsten, 2009] Horsten, L. (2009). Levity. *Mind*.

[Horsten, ta] Horsten, L. (t.a.). The Tarskian Turn: Deflationism and Axiomatic Truth.

[Horwich, 1992] Horwich, P. (1992). *Meaning*. Oxford University Press, Oxford.

[Horwich, 1998] Horwich, P. (1998). *Truth*. Clarendon Press, Oxford, second edition edition.

[Jech, 2002] Jech, T. J. (2002). *Set Theory*. Springer monographs in mathematics. Springer, Berlin, Heidelberg, New York, 3rd millenium edition.

[Kremer, 1988] Kremer, M. (1988). Kripke and the Logic of Truth. *Journal of Philosophical Logic*, 17:225 – 278.

[Kripke, 1975] Kripke, S. (1975). Outline of a Theory of Truth. *The Journal of Philosophy*, 72(19):690 – 716. Seventy-Second Annual Meeting Americal Philosophical Association.

[Kühnberger et al., 2005] Kühnberger, K.-U., Löwe, B., Möllerfeld, M., and Welch, P. (2005). Comparing Inductive and Circular Definitions: Parameters, Complexity and Games. *Studia Logica: An International Journal for Symbolic Logic*, 81:79–98.

[Lorenzen, 1955] Lorenzen, P. (1955). *Einführung in die operative Logik und Mathematik*. Springer, Berlin, Göttingen, Heidelberg.

[MacFarlane, 2000] MacFarlane, J. (2000). *What does it mean to say that logic is formal?* PhD thesis, University of Pittsburgh.

[McGee,] McGee, V. Field's Logic of Truth. *Philosophical Studies*, 147:421–432.

[McGee, 1994] McGee, V. (1994). Afterword: Truth and Paradox. In Harnish, R. M., editor, *Basic Topics in the Philosophy of Language*, pages 615–633. Harvester Wheatsheaf.

[Moriconi and Tesconi, 2008] Moriconi, E. and Tesconi, L. (2008). On Inversion Principles. *History and Philosophy of Logic*, 29(2):103 – 113.

[Prawitz, 1965] Prawitz, D. (1965). *Natural Deduction: A Proof-Theoretical Study*. Almqvist & Wiksell, Stockholm.

[Prawitz, 1994] Prawitz, D. (1994). *Mind*, 103(411):373–376.

[Priest, 2007a] Priest, G. (2007a). Revenge, Field and ZF. In Beall, J., editor, *Revenge of the Liar*, pages 225 – 233. Oxford University Press, Oxford.

[Priest, 2007b] Priest, G. (2007b). Spiking the Field-Artillery. In Beall, J., editor, *Deflationism and Paradox*, pages 41 – 52. Oxford University Press, Oxford.

[Priest, 2010] Priest, G. (2010). Hopes fade for saving truth. *Philosophy*, 85:109 – 140.

[Read, 2000] Read, S. (2000). Harmony and Autonomy in Classical Logic. *Journal of Philosophical Logic*, 29:124 – 154.

[Read, 2009] Read, S. (2009). Field's Paradox and its Medieval Solution. Talk presented to Logica.

[Read, 2010] Read, S. (2010). General-Elimination Harmony and the Meaning of the Logical Constants. *Journal of Philosophical Logic*.

[Reinhardt, 1986] Reinhardt, W. (1986). Some Remarks on Extending and Interpreting Teories with a Partial Predicate for Truth. *Journal of Philosophical Logic*, 15:219 – 251.

[Scott, 1975] Scott, D. (1975). Combinators and classes. In Bohm, C., editor, *λ-calculus and computer science*, Lecture Notes in Computer Science, page 1–26. Springer, Berlin.

[Simmons,] Simmons, K. *Universality and the Liar: An Essay on Truth and the Diagonal Argument*. Cambridge University Press, Cambridge, New York.

[Stephen Blamey, 2002] Stephen Blamey (2002). Partial Logic. In *Handbook of Philosophical Logic*, volume 5, pages 262 – 353. D.Reidel, second edition edition.

[Tarski, 1955] Tarski, A. (1955). A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5:285 – 309.

[Tarski, 1956] Tarski, A. (1956). The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics*. Hacket Publishing Company, 1983 edition.

[Troelstra and Schwichtenberg, 1996] Troelstra, A. S. and Schwichtenberg, H. (1996). *Basic Proof Theory*. Cambridge Tracts in Theoretical Computer Science 43. Cambridge University Press, Cambridge, second edition 2000 edition.

[Visser, 2004] Visser, A. (2004). Semantics and the Liar Paradox. In Gabay, D. and Günther, F., editors, *Handbook of Philosophical Logic*. Kluwer, 2nd edition.

[Welch, 2008] Welch, P. D. (2008). Ultimate truth vis-a-vis stable truth. *Review of Symbolic Logic*, 1:126–142.