Chapter 12
Norms for Theories of Reflexive Truth

Volker Halbach and Leon Horsten

Abstract In the past two decades we have witnessed a shift to axiomatic theories of truth. But in this tradition there has been a proliferation of truth theories. In this article we carry out a meta-theoretical reflection on the conditions that we should want axiomatic truth theories to satisfy.

12.1 Introduction

In the past two decades we have witnessed a shift from semantic—such as Kripke's or the revision theory—to axiomatic theories of truth. But in this line of research there has been a proliferation of truth theories. How should we adjudicate between them? [1]

We list some desiderata or norms for axiomatic theories of truth and explain how they can be used to discriminate between theories. To some extent these norms will also be useful for explaining what is driving the work on axiomatic theories of truth, as in many cases authors—including the authors of this article—have not been very explicit about their motivations, but rather concentrated on analysing the formal properties of the theories. So to a limited extent the norms we are going to list should be understood to be not only normative but also descriptive in the sense that they are intended to make explicit the norms that have be applied by various authors in the field.

Volker Halbach's work is part of the research project AH/H039791/1 that is supported by the Arts & Humanities Research Council.

[1] The title and some of the content of this paper is inspired by Giulia Terzian's (2012) PhD work on norms for theories of truth.

V. Halbach
New College, University of Oxford, Oxford, UK
e-mail: volker.halbach@philosophy.ox.ac.uk

L. Horsten
University of Bristol, Bristol, UK
e-mail: leon.horsten@bristol.ac.uk

© Springer Science+Business Media Dordrecht 2015

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_12

Of course the formal analysis of a theory will be helpful in assessing whether a theory satisfies the desiderata and the formal analysis cannot be separated from a philosophical judgement. We don't see the reflections on the norms as a stage that should precede formal work on the systems. Formal results have helped logicians to choose and formulate the norms, and the norms have motivated the formal work on theories. So we do not think of the norms as a *prima philosophia* that is conceptually prior to the logical analysis.

Some efforts have been made to carry out exercises similar to that of the present article: these include Sheard 1994 and Leitgeb 2007. But, for reasons that we hope to make clear, these efforts have remained less than fully satisfactory.

Let us start by listing some aspects of axiomatic truth theories that we require all axiomatic truth theories to accept wholesale. They will be treated as background assumptions in the sequel. This is done mainly to focus the discussion. We do not claim that they are unproblematic and cannot be challenged.

The aim is to explicate the meaning of the truth predicate without presupposing the distinction between object- and metalanguage. We will investigate systems in which the truth predicate is contained in the object language. So we mainly aim to unfold the notion of type-free or *reflexive* truth. The distinction between typed and type-free notions of truth is not unproblematic.[2] But all the theories that we are considering here prove the truth of sentences containing the truth predicate and are thus type-free in this sense.

We will treat the notion of truth as a primitive predicate T and reflect on how truth can and should be axiomatised. The treatment of truth as a primitive predicate in itself does not rule out the possibility that truth is a definable concept. Whether truth is definable or reducible by other means depends on the chosen axioms. However, only very weak theories will escape Tarski's theorem on the undefinability of truth. Hence the undefinability of truth in the case of all interesting truth theories will be a result that is arrived at, not a presupposition. So in contrast to semantic approaches and "substantial" theories, where the definability of truth has to be assumed from the outset, the axiomatic approach is neutral with respect to the question whether truth is definable or not.

It is commonly acknowledged that the axioms of truth should be studied in conjunction with axioms for the objects to which truth is ascribed. These may be sentence types or tokens, propositions, or still other objects. Many authors working on the axiomatic approach to truth think of truth as a predicate applying to (codes of) sentence types and we will follow them here without justifying this assumption.

The languages and theories we are interested in are formulated in first-order predicate logic. We hope that many of the results shed light on the use of the truth predicate in philosophical discussions, but we do not claim that our setting is the best starting point for an analysis of truth in natural language.

[2] For more on the distinction between typed and type-free theories of truth, (see Halbach 2011, Sect. 10).

The language in which the axiomatic theories are formulated will be taken to be $\mathcal{L}_T$, which is the language $\mathcal{L}_{PA}$ of first-order arithmetic expanded with the primitive truth predicate T. For again familiar reasons, we need in our axiomatic systems to be able to reason about finite sequences of symbols, or, equivalently, about finite sequences of numbers (conceived of as codes of symbols). We insist that all the axiomatic truth theories that we consider can prove all theorems of PA restricted to $\mathcal{L}_{PA}$. So we shall call PA (restricted to $\mathcal{L}_{PA}$) the *base theory*, and we shall call $\mathcal{L}_{PA}$ the *ground language*.[3]

Many of our comments will apply to many other base theories *mutatis mutandis*. Using theories weaker than PA will complicate the considerations in many cases. For instance, if a theory such as $I\Sigma_1$ employed, then it is far from being clear how the induction principle should be generalised to the expanded language with the truth predicate. We prefer to steer clear of these delicate issues in the present paper. Generally it is easier to apply our considerations to theories with unrestricted schemata like Zermelo–Fraenkel set theory. But also in this case some delicate issues arise, as is shown by Fujimoto 2012.

12.2 Imprecise and Contestable

We aim to formulate informative criteria that apply directly to axiomatic systems. Nonetheless, all our criteria are imprecise and fuzzy.

Fulfillment of our criteria is not going to be an all-or-nothing affair.[4] The criteria will be such that they can be satisfied to a lesser or to a greater extent. It should not be assumed that the degree of satisfying individual desiderata can be quantitatively measured: perhaps a partial ordering relation for truth theories is the best we can get.

Given that we are plagued by the semantic paradoxes, it will come as no surprise that the desiderata on our list will not be independent: they cannot all jointly be satisfied, as we will see. But given that fulfillment of the criteria will be a matter of degree, we may hope to be able to satisfy all desiderata to a reasonable degree. While the criteria are not independent, we do want a significant amount of independence between them.

Even if meaningful quantitative degrees could be obtained, evaluating the adequacy of an axiomatic theory of reflexive truth will not be a simple matter of adding degrees. We should not even assume that there is one "formula" that assigns the correct weight to each of the desiderata that we will propose.

We will not go as far as stating one single norm for truth theories. Some researchers believe that any list of norms for axiomatic truth theories should derive from a single purpose of truth, such as expressing generality, or from a single property such as "transparency". But we cannot see how the desiderata actually guiding the search for

[3] For a discussion of the ground language and the base theory see Halbach 2011 and Horsten 2011.

[4] This is also the case for most of Sheard's criteria, which we will discuss later. See Sheard 2002, p. 173.

attractive theories of truth can be derived from such a single purpose or property.[5] At least some non-triviality condition is driving the quest for theories of truth as well.

In fact we think the usability of truth as a device of generalisation is derived from more basic properties such as semantic ascent and perhaps also from compositionality. We do not deny that truth is a tool serving a specific purpose, but we do not think that whatever fits the purpose should be called truth. Rather we start with a predicate satisfying our or similar desiderata and then observe what purposes it can serve.

12.3 Five Desiderata

We shall now propose and discuss a series of desiderata for axiomatic theories of truth. We list them in no particular order of priority.

When authors appraise and advocate their theories of truth, they usually employ criteria of highly diverse kinds. Presumably the desiderata of the simplest kind are those that require the theory to prove certain theorems. For instance, the theory may be required to prove all T-sentences for all sentences from a certain class or the claim that truth commutes with conjunction.

Another kind of desideratum is formulated in more metatheoretic terms. For instance, the theory may be required to be ω-consistent (see Leitgeb 2007) or "symmetric" concerning its inner and outer logic, where the outer logic of a theory S is the set of S-provable sentences and the inner logic is the set of sentences φ with $S \vdash T\ulcorner\varphi\urcorner$ (see Halbach 1994; Leitgeb 2007).

Some desiderata formulated in metatheoretic terms are problematic for our purposes, because we aim to assess the strengths and weaknesses of a truth theory from the perspective of the base theory of that theory. This is especially important when we use our overall theory—this may be our most general theory containing set theory—, because then no standpoint external to this most comprehensive theory will be available to us.

This problem is not specific to truth theories but to overarching frameworks in general. For instance, postulating that our set theory, when used as a general foundations for mathematics, is consistent, cannot be used as a desideratum for the theories themselves: none of the theories can prove the existence of such a model (unless it is inconsistent). This does not imply that the desideratum cannot be used as a guide for extensions of the theory, but we will never be able to see from the standpoint of our best theory that it has a nice and well behaved model. What we can do in certain cases, is to prove in the base theory that if the base theory is consistent, then extending it with certain truth axioms will still yield a consistent theory.

However, in many cases the metatheoretic desiderata are provable in the axiomatic truth theory itself. For instance, the claim that the inner and outer logic coincide has

[5] In fact, we suspect that it has not been clarified what it means to "express generalisations". We are not really satisfied by Halbach's proposal (1999), for instance. Also transparency in itself does not seem to be the full story.

been advocated as a norm, and in most cases this claim that $T\ulcorner\varphi\urcorner$ is provable if and only if φ is provable is itself provable in the truth theory, or, in fact, already in weak arithmetical theories.

For truth theories even the postulate that there should be well behaved models can be reformulated and internalized to *some* extent in the object theory: We might expect that adding the truth axioms to a weaker base theory yields a theory with well behaved models. For instance, we might aim at a truth theory based on Zermelo–Fraenkel and consider whether adding the truth axioms to small fragments of ZF or Peano arithmetic results is a theory for which nice models can be constructed in ZF.

12.3.1 Coherence

Coherence is a notoriously underdetermined notion. Somehow the axioms and rules of a theory should be in harmony with each other. Coherence should not be understood in the way it is often understood in coherence theories of knowledge as supporting or even implying each other. We expect that the axioms and rules for truth do not clash with the base theory and the other axioms and rules for truth. But more may be said. If say the connective $\wedge$ and its interaction with the truth predicate is axiomatised in a certain way, then this would "cohere" with an analogous treatment of a connective like $\vee$.

If the truth theory contradicts the base theory, then the truth theory completely fails on the coherence norm. So if the truth theory proves the negation of an arithmetical theorem of Peano arithmetic, the theory does not cohere with the base theory. We adopt this requirement even for theories of truth in nonclassical logics, such as paraconsistent logics. Contradictions may not be lethal if the contradictions arise from the liar paradox and the theory proves both the liar sentence and its contradiction, but even the paraconsistent logicians will admit that a theory contradicting its own base theory is hopeless, because then the contradiction will be located in the non-semantic part of the theory that ought to be unproblematic. Perhaps exceptions may be made for extremely rich base theories that contain already problematic notions, but even then the addition of a truth predicate should not create any new contradictions in the ground language.

However, coherence is more than just plain consistency with the base theory. The axioms of a truth theory can also be incoherent in other ways.

Another example of a theory that clashes with the base theory is the Friedman–Sheard theory FS introduced under another name by Friedman and Sheard (1987) and further studied by Halbach (1994). The theory is internally consistent, but it is ω-inconsistent, that is, for some formula $\varphi(x)$ the system FS proves $\varphi(\overline{n})$ for each number n but it also proves $\exists x\ \neg\varphi(x)$. This fact can be proved in FS itself and thus FS is inconsistent with the uniform reflection principle $\forall x\ (\mathrm{Bew}_{FS}(\ulcorner\varphi(\dot{x})\urcorner) \rightarrow \varphi(x))$ for

FS.[6] It is also PA-provably inconsistent with the stronger global reflection principle stating that all closed theorems of FS are true. Hence the theory FS refutes its own soundness. We take this to be a form of incoherence.

Other authors such as Leitgeb (2007) have listed the existence of a standard interpretation as one of their norms for truth theories.

Certainly ω-consistency is an important requirement, but we do not see it as a very fundamental requirement: hardly anyone would have thought of imposing ω-consistency as a requirement before McGee (1985) proved the ω-inconsistency of a vast class of truth theories that looked otherwise very attractive. There are many different ways a theory can be incoherent with its base theory; ω-inconsistency is just one of them and McGee's proof showed that this form of incoherence can easily arise for theories of truth.

There may be other forms of incoherence with the base theory. For instance, if the induction schema cannot be extended to the language with the truth predicate on pain of inconsistency, then the theory in question is incoherent with the base theory. We are not aware of a natural theory of truth that violates this requirement, but if it should arise in the future, then the theory would have to be rejected as incoherent, even though nobody had imposed this requirement (except for us). The rejection of the theory would be justified, because of the incoherence with the base theory.

So far we have focused on the coherence of the axioms and rules for truth with the base theory. But a theory of truth should also be coherent in its truth-theoretic part. For instance, the truth theory may be internally inconsistent in the sense that the theory proves $T\ulcorner\varphi\urcorner$ for all sentences φ. Such a theory is not coherent and almost as bad as an inconsistent theory in classical logic.

We do not go into the discussion of the coherence of nonclassical systems. Paraconsistent logicians may want to say that their theories are coherent in some way. At least these theories can be non-trivial, but we are not quite sure whether this should be counted as evidence for coherence. The formulation of criteria for coherence may depend on the particular logic that is chosen.

12.3.2 Disquotation and Ascent

Deflationists do not tire of reminding us that truth is a disquotational device and a device for performing semantic ascent. Hence a sentence φ and the claim $T(\ulcorner\varphi\urcorner)$ should be equivalent in *some* sense. Some disquotation principles will also apply to formulae φ with free variables, but we will not consider them here.

A strong way of giving this slogan precise content is to desire that for a sentence φ of $\mathcal{L}_T$, φ and $T(\ulcorner\varphi\urcorner)$ can be substituted *salva demonstrabilitate* in any formula of $\mathcal{L}_T$. Field 2008 advocates this requirement under the label *transparency*. Others

[6] See Halbach 2011 for details.

focus on the Tarski-biconditionals, that is, equivalences of the form $T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$. Depending on the chosen logic, both requirements need not coincide.

The Tarski-biconditionals figure prominently in the theory of the paradoxes. Under fairly weak assumptions both the transparency principle and the full Tarski-biconditionals lead to inconsistency. Some philosophers, among them Tarski and more recently Burgess and Burgess 2011, do not flinch and accept the inconsistency view of truth. We have some sympathies with this view, but we do not pursue the project of giving an analysis of the truth predicate as found in everyday English. We see the quest for a good theory of truth as a revisionary enterprise, which involves the rejection of certain features of the truth predicate that may form part of a full analysis of the pretheoretic notion of truth.

There are many different ways to weaken the transparency principle and the Tarski-biconditionals to obtain a nontrivial theory or truth, and there are many different ways to classify these weakenings. We distinguish two possible ways to a non-trivial version of disquotation requirement: Either the class of instances of the T-schema is restricted or the connective $\leftrightarrow$ is replaced with some other (weaker and nonclassical) operator. Both methods can be combined.

On the first account, the set of instantiating sentences is restricted. Often the guiding principle seems to be to retain as many instances of the T-schema as possible. The qualification "as possible" has proved to be less tractable than expected. In particular, the restriction to *consistent* instances does not suffice, as McGee 1992 has shown by using a variant of Curry's paradox. So one might try to confine the set of permissible instances to those that do not prove any new theorems in the ground language. But Cieśliński 2007 showed that this policy does not fare much better than the bolder restriction to maximal consistent sets of instances.

In particular, using an argument reminiscent of Curry's paradox, McGee 1992 showed that in the presence of the diagonal lemma every sentence of $\mathcal{L}_T$ is equivalent to a Tarski-biconditional. Thus any sentence independent from the base theory can be decided using a consistent instance of the T-schema. The reasoning is very simple: Given a sentence φ one obtains a sentence γ_φ by the diagonal lemma:

$$\gamma_\varphi \leftrightarrow (T\ulcorner\gamma_\varphi\urcorner \leftrightarrow \varphi)$$

This equivalence logically implies that φ is equivalent to $T\ulcorner\gamma_\varphi\urcorner \leftrightarrow \gamma_\varphi$.

These equivalences are puzzling. If φ is the negation of a theorem of Peano arithmetic or the claim that PA is inconsistent, then the Tarski-biconditional $T\ulcorner\gamma_\varphi\urcorner \leftrightarrow \gamma_\varphi$ is incoherent with the base theory and ruled out by our first norm. In contrast, if φ is a theorem of PA, then $T\ulcorner\gamma_\varphi\urcorner \leftrightarrow \gamma_\varphi$ is provable at any rate without any further truth-theoretic assumptions and thus coherent. If φ is a sentence such as the consistency statement Con_{PA} of PA, the situation is more involved: It will be hard to argue that $T\ulcorner\gamma_{\mathrm{Con}}\urcorner_{PA} \leftrightarrow \gamma_{\mathrm{Con}_{PA}}$ is incoherent; after all it is equivalent to Con_{PA} and that sentence has been argued to be implicit in the acceptance of PA. In contrast, if φ is the consistency statement for a strong theory such as Zermelo–Fraenkel set theory the Tarski-biconditional $T\ulcorner\gamma_\varphi\urcorner \leftrightarrow \gamma_\varphi$ is at least not obviously coherent with any theory of truth with PA as base theory, because the consistency of such as strong theory is, so to speak, beyond the horizon of an arithmetical theory (assuming that it actually is consistent). So we suspect that there is no simple uniform policy concerning these McGee–equivalences as theorems of a truth theory and they seem to come in many degrees of implausibility, which makes any policy on restricting the disquotation schema tricky. We leave the treatment of these sentences to a future paper where also other restrictions on permissible instances of the T-schema will be discussed.

At any rate, there is no proposal to restrict the admissible instances of the T-schema that is commonly accepted. Even the restriction to T-free instances may be too liberal (Halbach 2006), although it is commonly rejected as too restrictive.

Hence the second option to render the Tarski-biconditionals nontrivial may look more attractive: The material biconditional between $T(\ulcorner\varphi\urcorner)$ and φ in the Tarski-biconditionals is replaced with some other operator or relation. The use of some binary predicate to this end has not become very popular. Expressing the claim that $T(\ulcorner\varphi\urcorner)$ and φ are materially equivalent in a single sentence in a straightforward way by means of a binary predicate expressing material equivalence requires a theory of this binary relation. But material equivalence between sentences φ and ψ is usually analysed as the truth of $\varphi \leftrightarrow \psi$. In general it does not seem attractive to base the theory of truth on a theory of some other relation.

More popular is the substitution of the biconditional by metatheoretic predicates. The transparency principle is an example; but it is too strong if classical logic is to be preserved. However, one can consistently demand that a truth theory S be closed under the inference rules $S \vdash \varphi \Rightarrow S \vdash T(\ulcorner\varphi\urcorner)$ (Necessitation) and $S \vdash T(\ulcorner\varphi\urcorner) \Rightarrow S \vdash \varphi$ (Co-necessitation). This is often paraphrased by saying that the outer logic of S coincides with the inner logic of S.

If classical logic is given up, the possibilities are multiplied. All kinds of so-called biconditionals may be used instead of the classical biconditional. Also metatheoretic principles that collapse into the Tarski-biconditionals in classical logic can be employed, even when the full Tarski-biconditionals are rejected. For instance, one can demand the truth theory S be closed under the inference rules $\varphi \Rightarrow T(\ulcorner\varphi\urcorner)$ and $T(\ulcorner\varphi\urcorner) \Rightarrow \varphi$. Of course in the presence of a deduction theorem, this demand yields the Tarski-biconditionals.

The satisfaction of the disquotation or semantic ascent criterion cannot even be ordered in a linear fashion. However, we do not think that the criterion can be abandoned, not even in favour of "compositional" axioms for truth.

In many cases the Tarski-biconditionals are strengthened by admitting free variables in the instantiating formulae. How this can be done depends on the chosen framework. In a purely arithmetic framework one can formalise the following statement for any formula $\varphi(x)$:

For any number n: $\varphi(\dot{n})$ is true iff $\varphi(n)$.

The dot above n indicates that the numeral of n is substituted for x in $\varphi(x)$. This can be expressed in a language with the resources of arithmetic as the substitution function is formally expressible in arithmetic.

There are versions with more than one free variable. In a more general setting, when not just numbers are admitted into the ontology, the use of a satisfaction predicate may be commendable. These versions of the Tarski-biconditionals with free variables are known as *uniform* or *parametrized* Tarski-biconditionals. We think that they flow from the disquotational feature of truth or satisfaction.

12.3.3 Compositionality

Davidson famously defended the view that truth is a compositional notion. In the present context with its very restricted ground language this means that the truth predicate commutes with the logical connectives and quantifiers such as $\wedge$, $\neg$, and $\forall$. This entails that truth of logically complicated sentences φ is determined by the truth value of logically "simpler" components of φ.

The compositional feature of truth does not contain its disquotational feature, at least not in an obvious way. Just demanding that truth commutes with all connectives will not suffice as long as the truth of atomic sentences is not regulated in any way. Usually "compositional" theories also contain the Tarski-biconditionals for T-free atomic sentences (often in a uniform, that is, parametrized version). This is the case for a Davidsonian conception. Once these are added, at least the Tarski-biconditionals for all T-free instances will be derivable under fairly general conditions.

Davidson and most of his followers have applied Tarski's solution to the paradoxes or have not cared much about the paradoxes in general. In particular, they have not said much about the truth of sentences that themselves contain the truth predicate and whether compositional semantics is possible for a language containing its own truth predicate. There are different ways of applying the compositionality requirement to sentences containing the truth predicate.

In the best case, truth should be expected to commute with the connectives and quantifiers independently of whether the sentences contain the truth predicate or not. Truth theories such as the Friedman–Sheard theory FS contain axioms stating this feature (see Halbach 2011).

However, the coherence requirement together with the disquotation desideratum may clash with full compositionality: FS contains all compositional axioms; it thus scores highly on compositionality. Moreover, it contains the rule version of the T-sentences, i.e., necessitation and co-necessitation. This yields an ω-inconsistency.

Therefore some truth theorists explicitly reject the full compositional axioms. In particular, many reject the axiom stating that truth commutes with negation while still retaining other compositional axioms. There are various motives for this restriction of compositionality. According to the view that rejects commutation of truth with negation, truth is fully compositional, but truth is a partial concept and does not apply to all sentences. In particular, the liar sentence may be said to be meaningless or the like and thus neither true nor false. Thus we should not expect the negation of the liar sentence to be true if the liar sentences is not true, because the negation of the liar sentence is as indeterminate as the liar sentence itself.

Thus in theories such as the Kripke–Feferman theory the compositional axioms are weakened to *positive* compositionality capturing the compositionality of a partial concept. The axioms for truth then no longer describe classical compositionality but rather the compositional principles of some nonclassical logic.

In the Kripke–Feferman theory this is combined with truth iteration axioms that have a strongly compositional flavour. Especially in Burgess' (2009) strengthening of Kripke–Feferman compositionality requirements seem to lead to a theory of *grounded* truth where the truth of each single truth depends on the truth of nonsemantical sentences, that is, sentences without the truth predicate. In this sense the groundedness requirement, which is occasionally seen as a desideratum for a truth theory, can perhaps be seen as a consequence of a strong version of the compositionality desideratum. However, the axiom that Burgess adds to the Kripke–Feferman principles can hardly be seen as expressing a form of compositionality.

Compositionality is not uncontroversial. In particular, supervaluationists argue that for languages containing vague expressions, truth does not distribute over all the connectives. We do not want to take a stance in this discussion. So the most we can claim here is that the compositional axioms should presumably be provable for the classical connectives and quantifiers as far as the vagueness-free fragments of first-order formalisations of vagueness-free fragments of natural language are concerned. Of course, if one goes beyond that, then there are serious worries concerning the possibility of a compositional semantics for natural languages.

12.3.4 *Sustaining Ordinary Reasoning*

Feferman famously rejected the possibility of withdrawing from full classical logic for $\mathcal{L}_T$ to partial logic as a way of avoiding the liar paradox, on the ground that "nothing like sustained ordinary reasoning" can be carried out in partial logic (Feferman 1984, p. 264). Thus a desideratum for axiomatic truth theories is that they should sustain ordinary reasoning.

Ordinary reasoning should be taken to include schematic mathematical or syntactic reasoning. So, for instance, in mathematics we are used to subjecting every predicate to mathematical induction. This means that axiomatic truth theories where the truth predicate is not allowed in the induction schema, do not receive a maximal score on this desideratum.

It is important to note that this demand extends not just to reasoning concerning sentences of the ground language but to sentences of the entire language $\mathcal{L}_T$. Our ordinary and mathematical reasoning is carried out in classical logic. Reasoning in intuitionistic logic does not come as natural to most of us, but it can be learned without too much difficulty. Reasoning in partial or in paraconsistent logic is a lot less natural still: it is very difficult to learn. Reasoning fluently in even more artificial logic, such as a logic in which certain structural rules are restricted (such as contraction, perhaps), might well be practically impossible.

It should also be mentioned that sustaining ordinary reasoning is not just a matter of the underlying logic and mathematics. If the truth laws themselves form a motley and scattered bunch, then even if the logical system containing it is fully classical, it will lose points on this desideratum. Also if for instance the inner logic is not classical, points are lost on this criterion. And this again underscores the fact that the criteria proposed here are not fully independent.

12.3.5 A Philosophical Account

Suppose that a theory of truth T is just given as a list of truth axioms in $\mathcal{L}_T$ added to a list of principles and rules of some logic. And suppose furthermore that T scores reasonably well against the norms discussed above: let us assume that T scores better against some of these desiderata than against others. Then some truth theorists would still find T thoroughly unsatisfactory as it stands. These truth theorists want in addition an explanation of why these norms are reasonable desiderata for a truth theory in the first place, and a justification of why it is acceptable that T does not satisfy each of the desiderata to the maximum extent. In sum, they request a philosophical account that justifies the norms, explains them, and ties them together. This request therefore is of a different nature than the other norms: it can be seen as a meta-norm.

The situation may be compared with that in set theory. If the naive theory of comprehension had proved useful (and consistent), then probably it could have passed more easily as a logical or almost logical principle that requires as little a philosophical story for its justification as the rules for the connectives.[7] But because a more sophisticated system such as Zermelo–Fraenkel with fairly complex axioms is required, philosophers felt that some philosophical account—such as Boolos' (1971) story of the cumulative hierarchy—is needed to motivate the axioms of set theory.

In the case of truth, the unrestricted Tarski-biconditionals are inconsistent under fairly general circumstances. The untyped axiomatic theories that have been proposed look far more sophisticated than the unrestricted Tarski-biconditionals. So one may ask whether there is a philosophical account analogous to that told by Boolos about set theory that can be used to motivate or perhaps even justify the truth-theoretic axioms, and that can explain, e.g., why not all the unrestricted Tarski-biconditionals are acceptable.

Many authors believe that a good axiomatic theory of reflexive truth should be embedded in a wider philosophical context and should be underpinned by a winning philosophical account that motivates the choice of the axiom. The philosophical story may comprise an account of how the content of the concept of truth is acquired, how new truths are in ordinary situations established on the basis of truths that have already been acquired (the revision theory of truth can be motivated such a story),

[7] The modality behind this counterfactual should probably an epistemic one. At any rate we do not claim that the inconsistency of comprehension is merely contingent. As with respect to a philosophical story about the rules for logical connectives, one might argue that some philosophical story is needed (that does not apply to 'tonk', for instance). But we would not classify this as a philosophical story that tells us something about the nature of conjunction.

how sentences of $\mathcal{L}_T$ that once were asserted are later withdrawn, how disagreements about propositions from $\mathcal{L}_T$ are resolved, or what purpose truth may serve. Classical "substantial" accounts of truth such as the utility view fall within this scope, but so does the story of the deflationists about truth as a device of generalisation. In this way, the axiomatic truth theory must somehow present a picture. It has to somehow expresses the main tenets of this philosophical account in a succinct and perspicuous way.

According to one such account, truth and falsehood are grounded in non-semantic facts: The truth of sentences ultimately supervenes on the truth of T-free sentences. This thesis is often seen as supported by Kripke's (1975) story about how the concept of truth is learned, and Kripke's minimal fixed point of the Strong Kleene scheme is regarded as one toy model of grounded truth. From iterated semantic ascent and compositionality it follows that many grounded truths of $\mathcal{L}_T$ ought to be included in the extension of the truth predicate. So including many grounded truths can be seen as a derived desideratum. This desideratum can be directly satisfied by a truth theory, by proving positive sentences that contain long iterations of the truth predicate T for instance. But it can also be indirectly met by containing natural interpretations of initial segments of the Tarskian compositional hierarchy, as described in Halbach 1995. Indeed, even from the point of reflexive truth it must be recognisable that the Tarskian hierarchy is fundamentally sound. Of course there is a limit to the length of the initial segment of the Tarski hierarchy that can be recovered by any axiomatic theory. The requirement that only grounded sentences should be classified as true or false cannot be regarded as derived from other desiderata. As mentioned earlier, it is implemented in a theory proposed by Burgess 2009.

Such a concomitant philosophical account or picture should not be confused with a mathematical model or class of mathematical models. In the axiomatic programme, models can at best have a heuristic use. Toy models can help us get a grip on an explanatory account. They can help us explore the structure of a philosophical account, which can then help us to formulate principles of truth. But that is all. In fact, as emphasised earlier, when we formulate our truth axioms for the strongest theories as base theory (set theory being an instance), then we will not even be able to establish the existence of models for the base theory to start from.

Here we do not a stance on the various philosophical accounts. We also do not exclude truth-theoretic pluralism that would admit incompatible truth theories as justified and underpinned by different philosophical accounts. For certain purposes we might be happy with one truth theory that may be better motivated by the actual use of truth in less theoretical contexts while another story is told about the purpose of the truth predicate in the philosophy of mathematics, and still another one about the purpose of the truth predicate in ethics.

The request for a philosophical account in the sense of this section may even be rejected altogether. One of the authors of this paper indeed denies the need for a philosophical account in this deeper sense. In some of the recent investigations into formal theories of truth (such as Friedman and Sheard 1987) this norm plays no role whatsoever. On the other hand, some research in the field has been motivated by desire to build axiomatic theories on the basis of a philosophical account.

12.4 Discussion

We claim that the desiderata that we have listed are more complete than rival lists that have hitherto been proposed. The hope is that all desiderata for theories of type-free truth derive from our list. But it can of course not be excluded that in the future desiderata for self-referential truth theories are discovered that are independent of the list that we propose here.

It is clear from our list of five desiderata that simple adding of marks on each dimension does not give a reliable judgement about the suitability of a theory of truth. For instance, if one is willing to accept a null score on coherence by giving up consistency altogether while classical logic is retained, then one can easily obtain maximal scores on most other dimensions and thus obtain a very high overall score. Yet most researchers would find such truth theories of little value.

Even though an individual researcher might find a very strong form of a desideratum (such as containing the unrestricted Tarski-biconditionals, or being completely classical) simply false, it seems unlikely that researchers who strongly support one of the criteria on the list—and there are many supporters for each of these criteria—are completely mistaken. So it seems not unreasonable to hold that a satisfactory theory of reflexive truth should do at least fairly well on each of the criteria on the list.

12.4.1 *Comparison with Sheard*

Sheard (2002) contains a list of maxims for truth theories. They are not intended specifically as desiderata for *axiomatic* truth theories. Nonetheless, it is worthwhile to compare them with our list.

Sheard's first maxim says that truth is an objective semantic concept. We agree on this point with Sheard, insofar as we understand it, and think that this point is captured by our background assumptions.

Sheard's second maxim says that provability preserves truth. What he means by this is that the axiomatic theory of truth has to be closed under the necessitation rule. This of course falls under our disquotation and ascent constraint. So we regard this as a specific gradation of one of our desiderata. A stronger version of Sheard's second maxim would hold that every truth theory S should be closed under the following rule of inference:

$$\frac{\begin{array}{c}\varphi\\ \vdots\\ \psi\end{array}}{T\ulcorner\varphi\urcorner \rightarrow \ulcorner\psi\urcorner}$$

In the presence of a deduction theorem, this rule is a consequence of our specific desiderata of compositionality and identity between internal and external logic.

Simplicity is Sheard's third maxim. We know from the literature in philosophy of science that simplicity is a theoretical virtue for scientific theories in general that is very difficult to spell out with any degree of precision. Simplicity should certainly not be equated with proof-theoretic weakness and, in particular, not with conservativity over the base theory. To some extent it is covered by coherence: the axioms of a theory should somehow hang together. Nonetheless, also in the case of truth theories we find it hard to determine fully in which way theories ought to be simple.

Sheard's fourth maxim is a difficult one: it says that often a "local truth analysis" is sufficient. The meaning of this is not completely clear to us, but it largely says that it is not necessarily required of a truth theory that it captures all aspects and uses of the truth predicate. In other words, this maxim is saying that it should not be a presupposition that all axiomatic truth theories can and should be compared to the same standard. It might be that one truth theory captures one use of the concept of truth very well, and another captures another very well, whilst no decent truth theory captures both at the same time. In our view, Sheard is onto something important here. This is why we have been explicit at the outset about what we expect our axiomatic truth theories to do: to give a decent account of reflexive uses of truth. This is what many (but not all!) contemporary axiomatic theories of truth are concerned with. And we do not want to claim that this is the sole cluster of uses of truth that one might be interested in capturing axiomatically. So Sheard's fourth maxim is one we have attempted to deal with in the preamble to our list.

As a fifth maxim, Sheard postulates the infinitary closure of the truth predicate. This seems a very specific requirement. It is captured by the formula

$$\forall x T\ulcorner\varphi(\dot{x})\urcorner \rightarrow T\ulcorner\forall x\varphi(x)\urcorner.$$

Of course this is a specific instance of our more general constraint of compositionality.

As a last and tentative maxim, Sheard contemplates requiring truth to be non-conservative. But he notes that this "almost always" follows from the previous maxim. He therefore in any case does not want to list it as a *fundamental* desideratum; we agree that it should be regarded as derived at most. But we take issue with Sheard's contention that it follows almost always from the previous maxim on Sheard's list (which is in our view also only a derived norm). It follows "almost always" from compositionality and the unrestricted presence of T in the induction scheme. But we have emphasized that there may be sound reasons for restricting the presence of truth in the induction scheme. So we cannot without qualification claim that non-conservativeness is even a derived norm for truth theories.

In general, our dissatisfaction with Sheard's list stems from the fact that his maxims are not fundamental enough.

12.4.2 Comparison with Leitgeb

Leitgeb compiles a list of norms that is significantly longer than ours and then goes on to discuss particular axiomatic truth theories as implementations of maximal consistent subsets of his list of norms. The norms on his list are mostly much closer to ours than Sheard's maxims. Let us take them in turn.

Leitgeb's first norm says that truth should be treated as a predicate, and his third norm says that truth should be treated as a type-free notion. Both of these belong to our background assumptions.

We skip Leitgeb's second norm for the nonce, and move on to number four on his list. This imperative says that good axiomatic truth theories should derive the unrestricted Tarski-biconditionals. This is a very strong form of our disquotation/ascent desideratum, but not necessarily the strongest one. Unrestricted substitution of φ and $T(\ulcorner\varphi\urcorner)$ in formulas of $\mathcal{L}_T$ is a stronger version, at least if certain weak underlying logics are assumed. It is one of the virtues of Field's theory of truth that it even satisfies the unrestricted substitutivity requirement of Field 2008.

Leitgeb's fifth norm is compositionality. As we have seen, this is also one of our norms.

The sixth norm on Leitgeb's list requires the existence of standard interpretations. We have discussed this norm above under our coherence norm.

Leitgeb's eighth norm requires the outer logic to be classical. This is a strong form of sustaining ordinary reasoning.

Norm seven on Leitgeb's list surprises us. It requires internal and external logic to coincide. It follows from norm four and norm eight. As intimated earlier, we see norm seven as a not-so-strong version of the disquotation / ascent desideratum. We suspect that Leitgeb lists norm seven because norm four, from which it can be derived in the presence of norm eight, is in effect by many researchers rejected on the ground of being excessively strong.

Let us, to conclude, turn to Leitgeb's norm two: it says that the truth theory should prove the truth of the background theory. In a weak form this can be derived from Leitgeb's requirement that the inner logic should coincide with the outer logic. But the uniform statement that all theorems of PA are true is of course a reflection principle from which nonconservativeness follows. We believe that Sheard is right in saying that this should not be a fundamental norm. If it is a norm at all, then it should derive from more fundamental considerations (involving considerations about the presence of truth in the induction scheme), and it is not so clear whether it follows without qualification.

In sum, we like Leitgeb's list better than Sheard's list. We agree with most of what is on its list, although we would regard many of his norms as strong instantiations of more general desiderata. But one desideratum is missing from Leitgeb's list just as it is missing from Sheard's list: we (or better: one of us!) want an axiomatic truth theory to capture a philosophical picture. And just as with Sheard's list, we object to the presence of non-conservativeness as an (almost) fundamental desideratum.

12.5 Applications

Supposing now that our list is a correct and complete list of fundamental norms for theories of reflexive truth: How can they then be applied?

It would be unreasonable to expect that by testing axiomatic theories of truth on our five dimensions, the question about which is (or are) the most satisfactory theory (theories) of reflexive truth can eventually be settled. One reason is that it is not completely clear how it can be measured how high a given theory scores on a given dimension. But a second, and perhaps more fundamental reason, is that different researchers will attach different weights to a given dimension. It is a familiar fact of the present situation that some researchers attach much value to ability to sustain ordinary reasoning, whereas other researchers attach much less weight to it. This is of course only to be expected, and it is an exact analogue of the situation concerning theoretical virtues of scientific theories. Some will put much stock on empirical adequacy, whereas others will be content to sacrifice some empirical accuracy to fruitfulness. So there are fundamental limitations on the use as a methodological tool of the norms discussed in this article.

This is not to say that judgements about weights attached to individual dimensions are not amenable to rational discussion. In fact, as we have remarked earlier, many researchers may be unwilling to accept that satisfying some particular criterion on the list to a high degree is desirable at all. Many researchers will take many of the unrestricted Tarski-biconditionals to be simply false, for instance. The most we can say is that everyone should agree that a satisfactory axiomatic system of reflexive truth should satisfy the norms on our list to a "reasonable" degree (except possibly the meta-requirement of giving a philosophical account).

There is a sense in which the norms discussed in this paper can be taken to demarcate an arena within which interesting formal theories of truth can be developed without stifling new research. Recall that Leitgeb has the requirement that inner logic coincides with external logic on his list but also the unrestricted Tarski-biconditionals and a demand for classicality in the outer logic. As we have remarked earlier, just classicality and the unrestricted Tarski-biconditionals would have met the case just as well. Proceeding in the way that he does, plays a large role in his identification of the main existing theories of reflexive truth as instantiations of the maximal consistent sublists of his list in a way that exhausts the maximal consistent sublists. His article is in a sense working towards this result. We do not aim at covering the maximal consistent degrees of satisfying our list with existing truth theories. We want to leave open the possibility and indeed hope that novel ways of having a mix of all the five norms to a reasonable degree will give rise to new interesting axiomatic theories of reflexive truth.

There is a related sense in which our methodological investigation might be of use for research 'on the ground'. Suppose we have an established and influential theory of reflexive truth that has a decidedly low score on some of the dimensions in the list, whilst having a high score on other dimensions on the list. Then we can attempt to modify the theory in such a way that we improve the score on the 'weak' dimension(s) whilst avoiding to push the scores on the 'strong' dimensions down significantly. Here is one brief example of how this can play out.

The Kripke–Feferman theory (KF), introduced by Feferman 1991 under the name Ref(PA), is one of the most popular axiomatic theories of reflexive truth. It does not score well on the disquotation / ascent dimension (since its inner logic does not coincide with its outer logic), whereas it does reasonably well on the other dimensions. The theory PKF proposed and investigated by Halbach and Horsten 2006 is close to KF, in that it is also an axiomatization of Kripke's truth theory. But PKF has unrestricted substitution of formulas φ by $T(\ulcorner\varphi\urcorner)$ and vice versa: thus it scores very well on the disquotation / ascent dimension. But of course there is a price to be payed. PKF is formulated in partial logic. So the question is whether the reduction of sustaining of ordinary reasoning is worth paying in exchange for improvement in on the disquotation dimension.

12.6 Conclusion

From the outside, it might look as if the field of axiomatic theories of reflexive truth is a methodologically strongly constrained enterprise. We hope that the present article has at least shown that this is far from the case. The methodological principles that are operative in this field are vague, variably adhered to, and pull in opposite directions. In this sense the field of axiomatic theories of reflexive truth does not differ from any other philosophical discipline.

References

Boolos, G. (1971). The iterative conception of set. *Journal of Philosophy*, 68, 215–231.

Burgess, A. G., & Burgess, J. P. (2011). *Truth*. Princeton: Princeton Foundations of Contemporary Philosophy. Princeton University Press.

Burgess, J. P. (2009). Friedman and the axiomatization of Kripke's theory of truth. Ohio State, 2009. paper delivered at the Ohio State University conference in honor of the 60th birthday of Harvey Friedman.

Cieśliński, C. (2007). Deflationism, conservativeness and maximality. *Journal of Philosophical Logic*, 36, 695–705.

Feferman, S. (1984). Towards useful type-free theories I. *Journal of Symbolic Logic*, 49, 75–111.

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

Field, H. (2008). *Saving truth from paradox*. Oxford: Oxford University Press.

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

Fujimoto, K. (2012). Classes and truths in set theory. *Annals of Pure and Applied Logic* (to appear).

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

Halbach, V. (1995). Tarski-hierarchies. *Erkenntnis*, 43, 339–367.

Halbach, V. (1999). Disquotationalism and infinite conjunctions. *Mind*, 108, 1–22.

Halbach, V. (2006). How not to state the T-sentences. *Analysis*, 66, 276–280 (Correction of printing error in vol. 67, 268).

Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.

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

Horsten, L. (2011). *The Tarskian turn: Deflationism and axiomatic truth*. Cambridge: MIT Press.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy,* 72, 690–712 (reprinted in 1984).

Leitgeb, H. (2007). What theories of truth should be like (but cannot be). In *Blackwell Philosophy Compass 2/2*, pages 276–290. Blackwell, 2007.

Martin, R. L. (ed.). (1984). *Recent essays on truth and the liar paradox*. Oxford, New York: Clarendon Press, Oxford University Press.

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

McGee, V. (1992). Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic,* 21, 235–241.

Sheard, M. (1994). A guide to truth predicates in the modern era. *Journal of Symbolic Logic,* 59, 1032–1054.

Sheard, M. (2002). Truth, probability, and naive criteria. In V. Halbach & L. Horsten (ed.), *Principles of truth*. Frankfurt a. M.: Dr. Hänsel-Hohenhausen.

Terzian, G. (2012). *Uncovering the norms of truth. A meta-theoretic inquiry*. PhD thesis, University of Bristol.