Philosophy Compass 2/2 (2007): 276–290, 10.1111/j.1747-9991.2007.00070.x

What Theories of Truth Should be Like (but Cannot be)

Hannes Leitgeb*
University of Bristol

Abstract

This article outlines what a formal theory of truth *should* be like, at least at first glance. As not all of the stated constraints can be satisfied at the same time, in view of notorious semantic paradoxes such as the Liar paradox, we consider the maximal consistent combinations of these desiderata and compare their relative advantages and disadvantages.

Formal theories of truth originated with Alfred Tarski ('Der Wahrheitsbegriff'; 'Semantic Conception of Truth'). They turned again into a hot topic of philosophical logic after Kripke, as well as Martin and Woodruff, had published their seminal pieces in 1975. In the meantime the number of contributions in this area has become legion; consequently, it is difficult to keep track of the aims that have shaped the field or may do so in the future.

The plan of this article is to outline what a formal theory of truth *should* be like, at least at first glance, i.e. the *prima facie* goals that we ought to reach for when we set up a theory of truth. We will see that not all of our intentions can be satisfied at the same time, so we are led to consider the maximal consistent combinations of these and compare their relative advantages and disadvantages.

Sometimes a distinction is made between *philosophical* theories of truth and *formal* or *logical* theories of truth. If such a distinction is granted at all, then this article should certainly be regarded as focusing on the latter. But ultimately every successful philosophical theory of truth has to stand the test of formalization and every successful formal theory of truth must be supported by philosophical argumentation; in this sense, the importance of the distinction should not be overestimated and also does not play a major role in the following considerations. The desiderata that we are going to deal with in the next section are simply desiderata for *theories of truth*. It is true that whether these desiderata are accepted or not depends, on the one hand, on one's philosophical background assumptions, and on the other, on the way in which these assumptions are formalized. But it would be wrong to think that either of these two sources of determination could ultimately be 'factored out' such that a 'merely' philosophical or a 'merely' formal theory of truth would be the outcome.

1. What Theories of Truth Should be Like . . .

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

There is almost unanimous agreement that truth is to be expressed by a predicate of the form 'is true' – briefly, '*Tr*' – and thus by a linguistic device that is applied to singular terms which are meant to denote the very objects that are true or untrue. For example, if '*Tr*' is a predicate of declarative sentences, then we want to concatenate it with proper names, definite descriptions or variables that refer to these sentences. In this way we are able to make claims such as:

(i) *Tr*('2 + 2 = 4')
(ii) *Tr*(the *x* such that *x* is the last sentence spoken by Caesar)
(iii) For all *x*, for all *y*: if *y* is the negation of *x*, then [*Tr*(*y*) if and only if not *Tr*(*x*)].

The main alternative would be to express truth by means of a sentential operator of the form 'it is true that' which is not applied to names but rather directly to sentences. While we can easily apply such an operator to sentences like '2 + 2 = 4' in order to form sentences such as 'it is true that 2 + 2 = 4', our (less trivial) examples (ii) and (iii) from above cause problems for this operator account. What if we do not know Caesar's last sentence and therefore cannot replace the definite description in (ii) by the sentence that it refers to? Furthermore, since 'for all *x* . . . it is true that *x* . . .' is not well-formed in standard first-order languages, how shall we relax our syntactic regimentation in order to harmonize truth operators with quantification? Using a truth predicate avoids all these complications from the start.

Once this issue is settled, we have to decide what the singular terms that we concatenate '*Tr*' with should refer to: (declarative) sentences, as above? Propositions? Utterances? Since a nice theory of sentences *qua* syntactic objects – the theory of syntax – is available, while the philosophical status of propositions and utterances is much less clear, we might opt for sentences just for the sake of simplicity and use the theory of syntax in order to describe which sentences there are, which syntactic properties they have, how they are built from other linguistic items, and so on. Indeed this is what we usually find in the literature on modern theories of truth, though sometimes it is not the theory of syntax itself that is employed as such a background theory, but instead the theory of arithmetic is used for this purpose. However, 'modulo' coding sentences effectively by natural numbers along the lines of Gödel's famous technique (now called 'Gödelization') this does not make any crucial difference. Indeed, Quine has taught us that sufficiently strong theories of syntax contain arithmetic up to encoding.

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

This is a rather uncontroversial point. Let us consider an example: Peano arithmetic is a purely mathematical first-order theory which includes sentences such as '2 + 2 = 4' or 'for all x, for all y: $x + y = y + x$' which are cast in arithmetical vocabulary. As long as we are only deriving sentences in this theory, we do not have to worry about truth predicates or other semantic concepts at all (indeed regular mathematicians certainly do not worry about semantic concepts). Now assume that we apply our theory of truth to Peano arithmetic: would it not be odd if arithmetical theorems such as '2 + 2 = 4' or 'for all x, for all y: $x + y = y + x$' could not be proved true from the combination of the two theories? If, for example, '*Tr*('2 + 2 = 4')' were not derivable, then the combined theory would claim that 2 + 2 = 4 without claiming that '2 + 2 = 4' is true; in such a case we should regard our theory of truth as incomplete since it omits an obvious truth. Even worse, if 'not *Tr*('2 + 2 = 4')' were derivable, such that any derivation of '*Tr*('2 + 2 = 4')' would lead to inconsistency, then we should consider our theory of truth to be falsified because it contradicts an obvious truth. Indeed, it should not only be possible to prove every single theorem of T true, but the general statement that says that every theorem of T is true, i.e. 'for all x, if x is provable in T then *Tr*(x)', should be derivable in a proper theory of truth. The same reasoning applies to cases where empirical theories – say, Newtonian mechanics – are used instead of mathematical theories. In a nutshell: a theory of truth should be designed in a way such that if truth is to be explained for the language of a certain theory T, then adding such a theory of truth to T should allow us to prove (the members of) T true, or otherwise this theory of truth would be either useless or flawed. (A related topic is whether a theory of truth should additionally be *conservative* over the theory T that it is supposed to extend, i.e. whether no new statements in the language of T should become derivable by adding a theory of truth to it: this has been debated hotly in the last few years in the context of deflationist theories of truth; we will not deal with this question any further, but see Shapiro; Ketland, 'Deflationism'; Tennant; Hyttinen and Sandu; *Mind* 114 for more details.)

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

If we agree that the sentence '2 + 2 = 4' is true, it is a minor step to admit that also the sentence '*Tr*('2 + 2 = 4')' is true. Accordingly, we want to claim that '*Tr*('*Tr*('2 + 2 = 4')')' is true, and so forth. This leads us to higher and higher levels of reflection, but there is nothing obviously wrong about this fact. Semantic ascents like these can be made more transparent by adding little indices to the truth predicate, thus saying '2 + 2 = 4', 'Tr_0('2 + 2 = 4')', 'Tr_1('Tr_0('2 + 2 = 4')')', . . . , yet there is no immediate necessity of doing so. Tarski's suggestion to regiment our formal languages in terms of a type-theoretic hierarchy of object languages, metalanguages, metametalanguages, . . . such that each of these language levels would have its 'own' truth predicate that is different from the truth predicates of the other levels,

was to some extent a response to the discovery of semantic paradoxes; we will return to this point later. But if the threat of paradoxes is disregarded for the moment, there are compelling reasons for using a simple *untyped* truth predicate. After all, this is what we do in natural language discourse (didn't you agree that '*Tr*('*Tr*('2 + 2 = 4')') is true?). Furthermore, as Kripke has made clear, there are perfectly fine applications of truth predicates in everyday language for which we would not even know what types should be assigned to them: for example, if 'the *x* such that *x* is the last sentence spoken by Caesar' refers to some sentence about Brutus and does not involve truth explicitly at all, then 'Tr_0' would have to be applied to it, but if Caesar had said something about the truth of such a sentence, then 'Tr_1' would have been the appropriate choice. Now assume we did not know what Caesar had said at the end of his life: it seems we would not be in the position to express the truth of the last sentence uttered by Caesar at all, which is strange in view of the fact that the *x* such that *x* is the last sentence spoken by Caesar is, presumably, perfectly "innocent" (at least as far as semantic considerations are concerned). So we should rather hold on to our original type-free truth predicate. Accordingly, nothing should stop us from applying the type-free predicate '*Tr*' to the name of a sentence that contains '*Tr*' as a predicate. If this is allowed for, we will say that the truth predicate is not subject to any type restrictions.

(d) T-biconditionals should be derivable unrestrictedly

Sentences of the form '*Tr*('A') if and only if A', in which the schematic letter 'A' is replaced by a declarative sentence of a given language, are called *T-biconditionals* for this language ('T' for truth, 'biconditional' because of the 'if and only if'); '*Tr*('A') if and only if A' itself is called the T-scheme. The famous '*Tr*('snow is white') if and only if snow is white' is the paradigm case example of a T-biconditional. Tarski was the first to notice the methodological importance of the T-scheme: his idea was to define truth, i.e. to state a definition of the form '*Tr*(*x*) if and only if . . .', and then to test the semantic adequacy of this definition by checking whether all T-biconditionals for the language for which truth is to be defined are derivable from it. If not, the definition is deficient. For example, if '*Tr*('snow is white') if and only if snow is white' were not derivable, then an important aspect of our understanding '*Tr*' as being applied to the sentence 'snow is white' would not be reflected by this definition. At best the definition would specify the actual meaning of '*Tr*' incompletely, in the worst case it might assign the wrong extension to '*Tr*'. The same seems to hold for all other sentences of the form '*Tr*('A') if and only if A' where 'A' is replaced by a sentence of the very language that we are interested in. Indeed, the derivability of all instances of the T-scheme from a definition of truth guarantees that the latter assigns the 'right' extension to '*Tr*', which at least seems to be a *necessary* condition for what a 'good' definition of truth must be like.

More recently, disquotationalists such as Field ('Deflationist Views') have argued that truth is governed by T-biconditionals *and nothing but T-biconditionals*. The latter do not even need to be derived from a definition of truth since they are self-sufficient as being the very axioms of truth: their triviality and a priori character is all that is needed to guarantee the truth predicate's function as a quasi-logical device. Even more recently, Field ('Semantic Paradoxes') and Beall ('Transparent Disquotationalism') have suggested that disquotationalists should simply demand the general intersubstitutivity of '*Tr*(*t*)' with the sentence denoted by '*t*' in every purely truth-functional context; the T-biconditionals follow from this postulate if additionally all sentences of the form 'A if and only if A' are accepted.

Independent of which theory of truth one prefers and on what grounds this preference is argued for, the derivability of all T-biconditionals for a given language is generally accepted as one of the prime desiderata for a theory of truth. What *is* disputed is the formal specification of the 'iff' that is used in T-biconditionals: the standard reading is of course material equivalence, but also reconstructions in terms of equivalence signs in non-classical logic (see the discussion of Field's theory below) or in terms of definitional equivalence (as in the Revision Theory of Truth; cf. Gupta and Belnap; Herzberger) have been suggested.

(e) Truth should be compositional

Suppose a sentence is built up from other sentences: whether or not this complex sentence is true seems to be determined solely by whether or not its syntactic constituent sentences are true and by the way the latter are put together. This phenomenon is usually subsumed under the umbrella term 'compositionality'; compositionality principles for truth, reference, meaning, and so forth, are among the fundamental principles of semantics. For example, why is it that we are in principle able to determine and understand the conditions under which a sentence is true independent of whether the sentence is new to us, as long as we are acquainted with the atomic particles and the logical structure of the sentence? The compositionality of truth yields an elegant explanation of this fact. Accordingly, Tarski used compositionality clauses in order to define inductively the truth or falsity of a complex sentence in terms of the truth or falsity of its logical parts. For example, (iii) from above is the compositionality principle that determines truth for the class of negation sentences, and analogous clauses can be introduced for conjunction sentences, disjunction sentences, and other syntactical categories (including quantified sentences, although the classes of universally and existentially quantified sentences demand some extra efforts which led Tarski to define a so-called satisfaction predicate first and then to base his definition of truth on this former definition). But compositionality by itself does not uniquely determine what a theory of truth looks like; for example, Kripke's Strong Kleene account of truth (see below) is compositional, however, while the truth value of a complex

sentence according to this theory is still determined solely by the truth values of its syntactic constituent sentences and by the manner the latter are composed, Kripke's theory allows for sentences which are *neither true nor false* or "undefined". As long as this absence of a classical truth value is also fixed compositionally – as it is the case with Kripke's theory – the theory still relies on compositionality as an essential semantic maxim, only the logic that underlies this theory is now different from classical logic.

(f) The theory should allow for standard interpretations

Speaking of the truth of a sentence without fixing an interpretation of the linguistic expressions within the sentence does not make much sense; without such an interpretation a sentence is not more than a sequence of meaningless signs arranged in accordance with a set of recursive rules. Usually, when we use a sentence we automatically assign an intended interpretation to it. For example, in describing an elementary mathematical fact by means of the sentence '2 + 2 = 4' we interpret '2' and '4' as referring to natural numbers, '+' as denoting the addition of natural numbers, and '=' as expressing the identity relation for natural numbers. When the arithmetician says 'for all x, for all y: $x + y = y + x$', then she understands 'x' and 'y' as ranging over natural numbers and only natural numbers; and so on. But if we are not careful enough, our intended interpretation of a formal language can be excluded by the theory that we formulate in this language and that we hold true. Standard examples are all first-order theories of the following kind: they are formulated in the language of arithmetic; they include an infinite sequence of sentences of the form $A[1]$, $A[2]$, $A[3]$, . . . ; at the same time they also include the sentence 'not for all x: A[x]'. It seems that any such theory would have to be inconsistent, but in fact it is an immediate consequence of the compactness theorem of first-order logic that such theories can in fact be consistent, i.e. no sentence of the form 'B and not B' is derivable from it. The appearance of inconsistency is effected by our intended reading of these sentences as having their standard arithmetical interpretation. It is this latter interpretation that is excluded by holding such a theory true: otherwise we would take $A[1]$, $A[2]$, $A[3]$, . . . to be saying that all natural numbers have the property A while 'not for all x: A[x]' would simultaneously express the contradictory opposite.

Returning to theories of truth again, when we use singular terms in order to express the truth or falsity of sentences as in our examples (i), (ii), (iii) from above, we intend these singular terms to refer to sentences, i.e. finite sequences of signs, and we intend the truth predicate to express a property of these sequences. A theory of truth should not exclude this standard interpretation, for otherwise the theory could not be understood as speaking about the very objects that it was designed to refer to. Put differently: a theory of truth does not only have to be consistent (of course it has to be!), it also should not mess up its intended ontological commitments. Tarski's theory of truth is the paradigm example of a theory that conforms to this

norm, but not every theory of truth in the field does. We will return to such theories below (see Leitgeb, 'Theories of Truth' for an overview).

(g) The outer logic and the inner logic should coincide

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

In a moment we will deal with the question of how to choose the logical systems on top of which we may introduce our theories of truth, but the topic of outer vs. inner logic is a different one: whatever reasons there might be for preferring one logic over another, if they apply to linguistic contexts outside the applications of truth predicates, why should they not equally apply to contexts within such applications? Every discrepancy between the outer and the inner logic of a theory of truth would indicate that our calling a sentence true somehow changes the logic that governs our understanding of this sentence. This is definitely questionable. Hence such discrepancies are – ceteris paribus – to be voted out. Note that this desideratum overlaps with the one in (b) but does not necessarily coincide with it: according to (b), if a theory of truth is applied to a mathematical or empirical theory, then the latter ought be provably true, which includes all logical truths that are formulated in the language of this theory. (g) ensures that this is the case for *all* logical truths whatsoever; if truth is to be explained for a language with truth predicate – which consequently would not be a language of a purely mathematical or empirical theory but would be partly semantic – then the same logical truths involving '*Tr*' should show up in the outer and the inner logic. Note that (g) is immediately implied by (d), as the instances of the unrestricted T-scheme can be used to import the laws of the outer logic into the inner one (given the logic of the 'if' in 'if and only if' allows for applications of *modus ponens*).

(h) The outer logic should be classical

If a theory is revised on rational grounds, the reasons for this change of theory are usually themselves consequences of another theory: one that is held constant and indeed presupposed in order to ground the revision of the other theory. If the *logical* axioms and rules of a theory are to be changed, it seems that *no* other theory could be held invariant, since every theory must include logic and thus will be affected by revising the latter. That is

the reason why some philosophers claimed that the selection of a logical system or, more generally, a linguistic framework, is of a merely pragmatic nature. As Carnap expressed it, there are no "morals" in logic; up to convention and convenience, anything goes. Recently Beall and Restall (2006) have argued for 'Logical Pluralism' on somewhat different grounds. According to a holistic account of theory change such as Quine's, any part of our scientific theories can be subject to revision as long as the changes increase the overall fitness of the theory in total; the logical parts of our theories are no exceptions to the rule, apart from their being more mentally and socially 'entrenched' than other parts.

What does this tell us about changes of the outer logic of theories of truth: is it justified to revise logical axioms and rules in view of some theory of truth that we would like to pursue and that would prove inconsistent or inelegant otherwise? Classical first-order logic is certainly the default choice for any selection among logical systems. It is presupposed by standard mathematics, by (at least) huge parts of science, and by much of philosophical reasoning. Therefore, every revision of it has to be supported by careful argumentation. For example, paraconsistent logicians such as Priest have argued in favor of a logic that allows for sentences to be *both true and false* at the same time without having the property that any sentence of the form 'A and not A' logically implies any other sentence. One (but not the only) argument for such a type of logic is its congeniality to a particularly simple theory of truth that can be based on it: the set of unrestricted T-biconditionals. (Some of these so-called 'dialetheist' theories of truth can be viewed as instantiating similar desiderata as Field's non-dialetheist theory which is discussed below.) However, suppose that two theories of truth are formally and philosophically equally useful and elegant, but while the one theory combines classical logic with a sophisticated axiom system for '*Tr*', the other theory compensates its obvious and plausible truth-theoretic axioms by deviating from classical logic in some sophisticated manner. It seems that in such a case the former theory should be preferred, if only because the principle of minimal mutilation tells us to be as conservative as possible, and fiddling with the semantic principles for '*Tr*' seems to be less "costly" than deforming our standard understanding of 'not', 'or', and the other logical constants. In this sense we should aim for the outer logic of truth to be classical, at least on a prima facie basis.

2. . . . But Cannot be

We have formulated eight norms that express what a theory of truth ought or ought not to be like – everything else being equal, and at first sight. The entries of this list are not meant to be independent of each other, nor is the list itself meant to be complete (a different though partially overlapping list of desiderata can be found in Sheard). In the best of all (epistemically) possible worlds, some theory of truth would satisfy all of these norms at the same time. Unfortunately, we do not inhabit such a world.

Let us take (a) and (b) for granted and let us assume also (c) for the moment. So we have to focus on (d) – (h). As was first observed by Tarski, (a) + (b) + (c) + (d) + (h) can give rise to inconsistency: Consider a first-order theory which conforms to these norms, such that truth is to be explained for the language in which this very theory is expressed. From the theory of syntax the existence of a so-called Liar sentence is derivable, i.e. a sentence L such that 'L if and only if not *Tr*('L')' follows from the syntactic axioms; up to provable equivalence, L says about itself that it is not true. If arithmetic is used as a theory of syntax, this is an immediate consequence of Gödel's famous Diagonalization lemma. By (d), every instance of the scheme for T-biconditionals for this language is derivable as well, whence '*Tr*('L') if and only if L' must be a theorem. Thus (h) allows us to apply classical reasoning in order to derive the equivalence '*Tr*('L') if and only if not *Tr*('L')', which is an inconsistency. Since (a) + (b) + (c) + (d) + (h) are inconsistent, (a) – (h) are so a fortiori (as Feferman 1984 has shown, the same result follows in intuitionistic logic).

Tarski's reaction to this consequence was to give up (c): the type restrictions that he introduced excluded all Liar-like sentences from being well-formed; accordingly, he did not any longer count such sentences as permissible substitutions for 'A' in the scheme for T-biconditionals. Indeed, Tarski was able to define '*Tr*' for languages without semantic expressions in a way such that the definition together with its formal background theory conformed to all postulates from above except for (c). However, as Kripke argued convincingly in his 'Outline of a Theory of Truth', the denial of (c) is both unnecessary and implausible. So let us rather take (a) – (c) for granted: how are we going to proceed from there? (This is actually an over-simplification: there are also theories of truth in which (c) is only dropped *partially*; see e.g. the contextualist approaches in Parsons; Burge; which were developed further by Glanzberg.)

Just as in the case of moral dilemmas, if a set of prima facie norms is not satisfiable simultaneously, the next best option is to search for *maximal* subsets that *can* be satisfied. Since (a) + (b) + (c) + (d) + (h) are inconsistent and (a) – (c) are considered given, every such maximal set may either contain (d) or (h) or none of the two, but definitely not both of them.

Let us first turn into the (h) direction, i.e. accepting classical logic: after dropping (d), the next candidate for a maximal satisfiable set of norms is (a) + (b) + (c) + (e) + (f) + (g) + (h). Can there be a theory of truth that conforms to this set? *No*. As McGee (1985) has shown, if compositionality principles are paired with classical logic both on the outside and the inside of applications of '*Tr*', then the resulting theory – though consistent – does not allow for standard interpretations (indeed it is of the same kind as the deviant theories that we have described under (f) above). This follows from the existence of a self-referential sentence which says that not all applications of the truth predicate (with quotation marks) to itself are true. Analogous results can be derived from the existence of an 'ungrounded' infinite sequence

of sentences such that each of the sentences in the sequence expresses that not all of its subsequent sentences are true (Yablo was the first to discover paradoxes of this sort (see 'Paradox without Self-Reference'); the question of whether such lists of sentences are self-referential or not has inspired lively discussions in the last few years: see Beall, 'Is Yablo's Paradox Non-Circular'; Leitgeb, 'What is a Self-Referential Sentence?'; Schlenker).

So let us take stock: since we have presupposed (a) – (c), since (d) is currently disregarded, (h) is assumed to be under investigation, and not all of (e) + (f) + (g) can be added to the former due to McGee's theorem, we are left with three further possibilities of maximal set of norms that might be satisfiable: (a) + (b) + (c) + (e) + (f) + (h), (a) + (b) + (c) + (e) + (g) + (h), and (a) + (b) + (c) + (f) + (g) + (h). It turns out that each of these sets can be realized by some theory of truth. Without claiming any sort of completeness whatsoever, we will briefly outline some representative instances of such theories.

Concerning (a) + (b) + (c) + (e) + (f) + (h): Kripke's theory of truth permits different kinds of formal specifications. One is in terms of a theory that is based on classical logic and which uses a fixed point construction in standard first-order set theory in order to define the extension of '*Tr*' for languages that include '*Tr*'; for example, if the language of arithmetic is extended by '*Tr*', then truth can be defined for it in this manner. While the construction is described in a classical language, it makes use of three-valued evaluations of the language for which truth is to be defined: at the initial evaluation stage, all sentences of the form '*Tr*(*t*)' receive the third truth value 'undefined'; at the next stage, '*Tr*(*t*)' gets the truth value 'true' ('false') if the sentence denoted by '*t*' had the truth value 'true' ('false') at the previous stage, and it is evaluated as 'undefined' otherwise; this is extended into the transfinite until a stable evaluation is reached (which can be proven to happen at some transfinite ordinal number). Sentences such as '2 + 2 = 4', '*Tr*('2 + 2 = 4')', '*Tr*('*Tr*('2 + 2 = 4')')', . . . , which are evaluated as true or false at the final stage, are called 'grounded'; sentences such as the Liar turn out to be non-grounded and hence truth-valueless.

Kripke's primary suggestion is to employ three-valued valuations according to the so-called Strong Kleene scheme (see Kripke). This has the effect that while the outer logic of this theory is classical, the inner logic is a system of partial logic; whence the outer and the inner logic do not coincide (Halbach and Horsten have recently studied an axiomatic version of the Kripke's Strong Kleene theory in which both the outer and the inner logic are partial; the corresponding system may be regarded as a subsystem of Field's below). On the other hand, truth is compositional according to this theory, the theory permits standard interpretations of its vocabulary, and if, for example, Peano arithmetic is available in the background, then it is provable that all theorems of Peano arithmetic are true. Yablo ('Grounding, Dependence, and Paradox') has reconstructed the same theory in terms of dependency relations: the true sentences as being given by Kripke's Strong Kleene

construction turn out to be those sentences whose truth values depend on the part of the language in which the truth predicate is *not* used. Feferman ('Reflecting on Incompleteness') has shown that instead of defining truth explicitly by set-theoretic methods, an elegant axiomatic system can be developed which has the same merits (or shortcomings) as Kripke's theory but which does not rely on set theory. Kripke's construction may be regarded as yielding the standard model for Feferman's theory. Maudlin is – or rather can be interpreted as (cf. Field, 'Maudlin's *Truth and Paradox*') – a recent defense of this Kripke-Feferman account of truth. Recently, Feferman (*Alfred Tarski Lectures*) has developed the (a) + (b) + (c) + (e) + (f) + (h) option into a different direction: advancing a line of reasoning pursued also by Martin ('Category Solution'), Martin and Woodruff, and McDonald, truth and falsity are regarded as properties only of *meaningful* sentences; the standard model of the resulting axiomatic system of truth and meaningfulness is given by the minimal Kripkean fixed point for the three-valued Weak Kleene scheme according to which, for example, the disjunction of the true sentence '2 + 2 = 4' and a meaningless sentence such as the Liar sentence is itself meaningless (the same disjunction would be true in the Strong Kleene scheme).

Concerning (a) + (b) + (c) + (e) + (g) + (h): If the outer logic of a theory of truth is classical and the inner logic is identical to the outer logic, then the former must be classical as well. A corresponding axiom system has been devised by McGee ('How Truthlike Can a Predicate be?') as well as Friedman and Sheard (and was studied further by Halbach): The language of the theory could be the first-order language of arithmetic extended by '*Tr*' again. The theory itself includes compositionality postulates for each of the classical connectives. The collapse of outer and inner logic is forced by a derivation rule which states that once A is proved then *Tr*('A') can be proved as well. Thus every law of the outer logic can be transferred into the context of the truth predicate and the same holds for all theorems of, say, Peano arithmetic, if the latter is presupposed axiomatically. As we have explained above, such a theory does not allow for standard interpretations; however, its consistency can be proved by means of a revision-theoretic construction as suggested by Gupta and Belnap, and Herzberger, and it can also be shown that the theory does not contain any false arithmetical sentences.

Concerning (a) + (b) + (c) + (f) + (g) + (h): Kripke noted that apart from the Strong Kleene scheme other evaluation functions can be used in order to support a fixed point construction for truth. A particularly interesting choice is the so-called Supervaluation scheme, which goes back to some of van Fraassen's ideas on free logic. Definitions by supervaluation allow for classical logic both as the outer and the inner logic of a theory of truth while at the same time standard interpretations of such theories are not excluded. On the other hand, truth is no longer compositional: for example, for every Liar sentence L the sentence '*Tr*('L or not L')' can be derived since 'L or not L' is a logical truth, however both 'not *Tr*('L')' and 'not *Tr*('not L')' are

derivable, too; so the compositionality principle '*Tr*('A or B') iff *Tr*('A') or *Tr*('B')' does not hold for all A, B of, say, the language of arithmetic extended by '*Tr*' again. As in the case of Kripke's Strong Kleene theory of truth, the Supervaluation theory can also be reconstructed on the basis of a semantic dependence relation (this was shown by Leitgeb's 'What Truth Depends on' for a fragment of this theory and can be proved for the total theory by related methods, as pointed out in the last section of Leitgeb, 'Towards a Logic'). Once again, true sentences prove to be grounded in the sense that their truth depends directly or indirectly just on the truth of sentences without truth predicate. The main difference between dependency according to the Strong Kleene theory and dependency according to Supervaluation is that the former is compositional while the latter is not. Cantini developed an axiomatic system which stands to the Supervaluation construction as Feferman's theory stands to the Strong Kleene model.

Let us now return to our observation that every maximal satisfiable subset of our set of eight norms may either contain (d) or (h) or neither but not both of them. What is going to happen if we opt for (d) – unrestricted T-biconditionals – rather than (h)? The corresponding 'optimal' candidate for a maximal satisfiable set would thus be (a) + (b) + (c) + (d) + (e) + (f) + (g) and a revision of classical logic would be the price to be paid. As it turns out, recent advances in the area of truth theories indicate that this set of norms is actually satisfiable:

Concerning (a) + (b) + (c) + (d) + (e) + (f) + (g): Field ('Revenge-Immune Solution') adds a new conditional sign '=>' to the logical constants of first-order languages. The logical laws that govern '=>' are weaker than the corresponding principles for material implication; for example, importation – 'A => (B => C)' logically implies '(A and B) => C' – fails. All of the other logical signs are axiomatized in agreement with the semantic rules of three-valued Strong Kleene logic. T-biconditionals, which are now formulated by means of the new conditional sign, are assumed unrestrictedly, whence if the theory is added to mathematical or empirical theories T that are expressed in the language for which truth is to be explained, all sentences that can be derived in these theories are provably true (though additional efforts have to be made to guarantee the derivability of the universally quantified sentence that expresses that all theorems of T are true). Accepting all T-biconditionals also implies that the inner logic must coincide with the outer logic and that neither is classical. Finally, the theory allows for standard interpretations, which Field is able to show by an ingenious blend of Kripkean minimal fixed points and the revision semantics. The only remaining question is whether this theory of truth is also compositional. As far as the Strong Kleene connectives are concerned, this is certainly the case. At first glance it seems obvious that truth must also be compositional concerning '=>', since every instance of '*Tr*('A => B')' <=> (*Tr*('A') => *Tr*('B'))' follows from the T-biconditionals for 'A => B', 'A', 'B', and from logical rules. This guarantees that the compositionality of

truth with respect to conditionals will be inherited from the compositionality of '=>', *if the latter obeys compositionality itself*. In contrast to the connectives of classical logic and three-valued Strong Kleene logic this is not obvious, therefore the compositionality of truth according to Field's theory is not to be regarded settled yet, nor is Field's claim that his theory is immune against the 'revenge' of some 'Super-Liar' paradox. (Beall's *Revenge of the Liar* contains Field's most developed version of his theory as stated in his 'Solving the Paradoxes, Escaping Revenge'; in the same volume, various replies and criticisms of Field's theory can be found.)

So where does this leave us? Can we rank our eight postulates in a way that would permit us to impose some additional 'order of acceptance' on the class of our maximal satisfiable sets? Should we take into account to what 'degree' a postulate is satisfied or dissatisfied? (For example, Field's theory still allows for classical logic in the '*Tr*'-free part of the language; on the other hand, the Supervaluation theory implies a restricted class of 'grounded' T-biconditionals.) Which other norms do exist that govern our understanding of truth? Do some of the norms from our list above have to be revised or dropped because their validity is actually restricted to theories of truth for languages *without* a truth predicate? Should we extend our semantic considerations of truth to semantic-*pragmatic* ones as it proved useful in other areas of philosophy? (Contextual theories of truth such as Glanzberg seem to point into such a direction.) Fortunately, there are a lot of questions left for future work on formal theories of truth. Despite some cooling down in the last fifteen years, the windchill factor of this area of philosophical logic is still far from 'freezing'.

(Concluding remark: though somewhat outdated, Martin's *Recent Essays on Truth and the Liar Paradox* is still the most important collection of articles in this area. McGee's *Truth, Vagueness, and Paradox* as well as Gupta and Belnap's *Revision Theory of Truth* are the corresponding best monographs and give an excellent introduction to this field of research.)

Short Biography

Hannes Leitgeb completed a Masters (1997) and a Ph.D. degree (1998) in mathematics and a Ph.D. degree (2001) in philosophy, each at the University of Salzburg where he later also worked as an Assistant Professor at the Department of Philosophy. In 2003 he received an Erwin-Schrödinger Fellowship from the Austrian Research Fund on the basis of which he did research at the Department of Philosophy and the CSLI at Stanford University. In 2005 he took up a joint position as a Reader at the Departments of Philosophy and Mathematics in Bristol. His research interests are in logic, epistemology, philosophy of mathematics, philosophy of language, and cognitive science. He has authored papers in these areas for: *Journal of Philosophical Logic*, *Synthese*, *Analysis*, *Philosophia Mathematica*, *Erkenntnis*, *Journal of Logic, Language and Information*, *Studia Logica*, *Notre Dame*

Journal of Formal Logic, *Topoi*, *Logique et Analyse*, and *Artificial Intelligence*. His monograph *Inference on the Low Level. An Investigation into Deduction, Nonmonotonic Reasoning, and the Philosophy of Cognition*, which appeared in 2004 in the Kluwer/Springer Applied Logic series, develops a logical and epistemological theory of justified inference within artificial neural networks. At present, he is trying to resurrect Rudolf Carnap's classic *The Logical Structure of the World*.

Note

* Correspondence: Department of Philosophy, University of Bristol, 9 Woodland Road, Bristol, Clifton BS8 1TB, UK. Email: Hannes.Leitgeb@bristol.ac.uk.

Works Cited

Beall, J. C. 'Is Yablo's Paradox Non-Circular'. *Analysis* 61 (2001): 176–87.
—— (ed.). *Revenge of the Liar*. Oxford: Oxford UP, forthcoming.
——. 'Transparent Disquotationalism'. *Deflationism and Paradox*. Eds. J. C. Beall and B. Armour-Garb. Oxford: Oxford UP, 2005. 7–22.
—— and G. Restall. *Logical Pluralism*. Oxford: Oxford UP, 2006.
Burge, T. 'Semantical Paradox'. *Journal of Philosophy* 76 (1979): 169–98. Reprinted in R. L. Martin (ed.). *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford UP, 1984.
Cantini, A. 'A Theory of Formal Truth Arithmetically Equivalent to ID_1'. *Journal of Symbolic Logic* 55 (1990): 244–59.
Feferman, S. *The Alfred Tarski Lectures*, Lecture 1: 'Truth Unbound'. Berkeley, CA, April 2006.
——. 'Reflecting on Incompleteness'. *Journal of Symbolic Logic* 56 (1991): 1–49.
——. 'Towards Useful Type-Free Theories. I'. *Journal of Symbolic Logic* 49 (1984): 75–111. Reprinted in R. L. Martin (ed.). *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford UP, 1984.
Field, H. 'Deflationist Views of Meaning and Content'. *Mind* 103 (1994): 249–85.
——. 'Maudlin's *Truth and Paradox*'. *Philosophy and Phenomenological Research*. Forthcoming.
——. 'A Revenge-Immune Solution to the Semantic Paradoxes'. *Journal of Philosophical Logic* 32 (2003): 139–77.
——. 'The Semantic Paradoxes and the Paradoxes of Vagueness'. *Liars and Heaps*. Eds. J. C. Beall and M. Glanzberg. Oxford: Oxford UP, 2003. 262–311.
Friedman, H., and M. Sheard. 'An Axiomatic Approach to Self-Referential Truth'. *Annals of Pure and Applied Logic* 33 (1987): 1–21.
Glanzberg, M. 'A Contextual-Hierarchical Approach to Truth and the Liar Paradox'. *Journal of Philosophical Logic* 33 (2004): 27–88.
Gupta, A. and N. Belnap. *The Revision Theory of Truth*. Cambridge: Cambridge UP, 1993.
Halbach, V. 'A System of Complete and Consistent Truth'. *Notre Dame Journal of Formal Logic* 35 (1994): 311–27.
—— and L. Horsten. 'Axiomatizing Kripke's Theory of Truth'. *Journal of Symbolic Logic* 71 (2006): 677–712.
Herzberger, H. G. 'Notes on Naïve Semantics'. *Journal of Philosophical Logic* 11 (1982): 61–102. Reprinted in R. L. Martin (ed.). *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford UP, 1984.
Hyttinen, T. and G. Sandu. 'Deflationism and Arithmetical Truth'. *Dialectica* 58 (2004): 413–26.
Ketland, J. 'Deflationism and Tarski's Paradise'. *Mind* 108 (1999): 69–94.
——. 'Yablo's Paradox and Omega-Inconsistency'. *Synthese* 145 (2005): 295–302.
Kripke, S. 'Outline of a Theory of Truth'. *Journal of Philosophy* 72 (1975): 690–716.
Leitgeb, H. 'Theories of Truth which have No Standard Models'. *Studia Logica* 68 (2001): 69–87.
——. 'Towards a Logic of Type-Free Modality and Truth'. *Logic Colloquium 05*. Ed. C. Dimitracopoulos. Lecture Notes in Logic, Association for Symbolic Logic, forthcoming.

——. 'What is a Self-Referential Sentence? Critical Remarks on the Alleged (Non-)Circularity of Yablo's Paradox'. *Logique et Analyse* 177–78 (2002): 3–14.
——. 'What Truth Depends on'. *Journal of Philosophical Logic* 34 (2005): 155–92.
McDonald, B. Edison. 'On Meaningfulness and Truth'. *Journal of Philosophical Logic* 29 (2000): 433–82.
McGee, V. 'How Truthlike Can a Predicate be? A Negative Result'. *Journal of Philosophical Logic* 14 (1985): 399–410.
——. *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*. Indianapolis, IN: Hackett, 1990.
Martin, R. L. 'A Category Solution to the Liar'. *The Paradox of the Liar*. Ed. R. L. Martin. New Haven, CT: Yale UP, 1970. 91–112.
—— (ed.). *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford UP, 1984.
—— and P. W. Woodruff. 'On Representing "True-in-L" in L'. *Philosophia* 5 (1975): 213–17. Reprinted in R. L. Martin (ed.). *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford UP, 1984.
Maudlin, T. *Truth and Paradox*. Oxford: Oxford UP, 2004.
Parsons, C. 'The Liar Paradox'. *Journal of Philosophical Logic* 3 (1974): 381–412. Reprinted in R. L. Martin (ed.). *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford UP, 1984.
Priest, G. 'Paraconsistent Logic'. *Handbook of Philosophical Logic*. 2nd ed. Vol. 6. Eds. D. Gabbay and F. Guenthner. Dordrecht: Kluwer, 2002. 287–393.
Quine, W. V. 'Concatenation as a Basis for Arithmetic'. *Journal of Symbolic Logic* 11 (1946): 105–14. Reprinted in W. V. Quine, *Selected Logic Papers*, enlarged ed. Cambridge, MA: Harvard UP, 1995.
Schlenker, P. 'The Elimination of Self-Reference (Generalized Yablo-Series and the Theory of Truth)'. *Journal of Philosophical Logic*. Forthcoming.
Shapiro, S. 'Proof and Truth: Through Thick and Thin'. *Journal of Philosophy* 95 (1998): 493–521.
Sheard, M. 'Truth, Provability, and Naive Criteria'. *Principles of Truth*. Eds. V. Halbach and L. Horsten. Frankfurt a.M.: Hänsel-Hohenhausen, 2002. 169–81.
Tarski, A. 'The Semantic Conception of Truth and the Foundations of Semantics'. *Philosophy and Phenomenological Research* 4 (1944): 341–76.
——. 'Der Wahrheitsbegriff in den formalisierten Sprachen'. *Studia Philosophica* 1 (1935): 261–405. Trans. and reprinted in A. Tarski, *Logic, Semantics, Metamathematics*. Oxford: Clarendon Press, 1956.
Tennant, N. 'Deflationism and the Gödel-Phenomena'. *Mind* 111 (2002): 551–82.
Yablo, S. 'Grounding, Dependence, and Paradox'. *Journal of Philosophical Logic* 11 (1982): 117–37.
——. 'Paradox without Self-Reference'. *Analysis* 53 (1993): 251–2.

