Review of Leon Horsten, *The Tarskian Turn*

In his new book *The Tarskian Turn* (MIT Press: Cambridge, MA, 2011; xii + 165 pages), Leon Horsten suggests that Tarski's pioneering work on truth in the 1930s can be seen as providing 'an emancipation of truth theory from traditional philosophy' (15). The reason is that Tarski did not ask 'Pilate's question' (11), viz., 'What is the essence of truth?' but instead attempted to answer the more minimal question, 'How does truth function?' by giving a formal theory which makes precise predictions about its behavior – i.e., about the manner in which truth is distributed over the truth-bearers. Horsten takes this shift in focus to be analogous to that which took place when Galileo and Newton explicitly avoided asking after the metaphysical nature of forces, instead giving simple descriptions of the laws concerning them (14-19); and it is this change in outlook regarding the investigation of truth which he regards as constituting the 'Tarskian turn' of the title. In short, Horsten regards Tarski as the first deflationist about truth.

The stated aim of the book is to 'bridge the gap between the philosophical and logical literature about the concept of truth' (xi). More specifically, beginning with Tarski's own work, a number of different logical (or perhaps mathematical) theories of truth have been developed, and Horsten is interested in the questions: (a) which of these theories is correct; and also (b) whether that best theory is compatible with philosophical deflationism. In the end he defends a theory for which he himself provided the proof theory (in joint work with Volker Halbach); and he argues that the conception of truth it articulates is indeed deflationist in character.

The first part of the book (chapters 1-3) is introductory. Here, and throughout, Horsten does an excellent job of presenting the required technical material in 'elementary terms' (xi), thus making it 'accessible to the average analytical philosopher' (xi). He does not start quite from scratch – he notes that it would be 'beneficial if the reader has taken an intermediate logic course' (6) in which a 'standard detailed proof of Gödel's completeness and incompleteness theorems' (7) has been given – but he does at least state, and in some cases sketch proofs of, these and other central metatheorems (e.g. the diagonal lemma) *en route* to proving Tarski's undefinability theorem, which says that 'no sufficiently expressive language can define its own truth predicate' (27). This crucial early theorem has immediate consequences for formal theories of truth. As Horsten notes (20), truth theories are of two kinds: *syntactic* (axiomatic, or proof-theoretic) and *semantic* (definitional, or model-theoretic). Tarski's theorem tells us that we cannot have a semantic theory of truth for our own language (i.e. the language in which the theory is given), a fact which leads Horsten to favour syntactic theories of truth (20-21). But the undefinability theorem also has consequences for syntactic theories: in particular, given certain standard assumptions, '[i]t implies that no consistent truth theory implies all the Tarski-biconditionals' (27) – i.e., all the instances of the schema

(*T*) The sentence ⌜φ⌝ is true if, and only if, φ.

These biconditionals, however, can be motivated by appeal to the *disquotational intuition* – according to which one who is 'willing to hypothetically assume or outright assert that φ' (17) ought (rationally) also to be willing to assume or assert that ⌜φ⌝ is true, and *vice versa* – together with the *deduction theorem*

of classical logic, 'which says that a sentence ψ is derivable… from a sentence φ if and only if the sentence $\varphi \rightarrow \psi$ is a theorem' (18). Tarski's theorem shows, however, that we must abandon the *naïve theory* of truth (*NT*), which simply takes as axioms all of the instances of the schema *T*, and it sets the agenda for further work on truth: the aim is to construct formal theories that are *philosophically sound*, proving no untruths about truth, but which are also as close to being *truth-theoretically complete* as is consistent with Gödelian incompleteness.

In the second part of the book (chapters 4-7), *typed* theories of truth are discussed, in which iterated truth ascriptions (of the form ⌜⌜φ⌝ is true⌝ is true) can be proved only if certain 'hierarchy constraints are satisfied' (103) by the truth predicates involved. There is a great deal that is of philosophical interest here and in what follows. Deflationism is the view that the essence of truth does not outstrip the conception that we have of it; and many deflationists (e.g. Horwich) have taken the disquotational intuition as, in effect, both constitutive and exhaustive of that conception. Thus, they have been impressed by the thought that we should simply collect together as many instances of the schema *T* as possible (without inconsistency), taking the result as our formal theory of truth. But McGee has shown that there are many maximal consistent collections of this kind; accordingly, some non-arbitrary means of deciding which of the Tarski-biconditionals should constitute one's truth theory must be sought. One such way is to rule out the application of 'true' to any sentence already containing that word: the result (endorsed by Quine) is the disquotational theory of truth (*DT*). But *DT* is very weak: it does not prove, for instance, that a universally quantified claim is true if and only if all of its instances are; and similarly for the other logical connectives. Horsten therefore embraces in addition the *compositionality intuition* 'that the truth value of a complex sentence is determined by the truth values of its component parts' (72). This leads him to favour the compositional theory of truth (*TC*) over *DT* (as, arguably, did Davidson); and as this theory enables us to prove new theorems (e.g. the consistency of Peano Arithmetic) which our base theory (Peano Arithmetic itself, serving as a surrogate for a theory of syntax) did not prove, Horsten rejects the claim, endorsed by some, that a theory of truth is deflationist only if it is mathematically conservative. The discussion of this, and related issues, in chapter 7, is fascinating – especially, to my mind, the material surrounding the theories *TC*- and *PT*-, and the notion of relative interpretability – and Horsten suggests that the 'weary intellectual traveller' (8) can readily set the book aside at this point.

But such a reader would, I believe, miss out. In the third and final part of the book (chapters 8-10), Horsten discusses *untyped* (or reflexive) theories in which truth iterations involving a single, univocal truth predicate can be proved. He begins with the Friedman-Sheard theory of truth (*FS*), which axiomatizes the (semantic) revision theory of Gupta and Belnap, and which sacrifices the full generality of the disquotational intuition (by rejecting the equivalence of φ itself with the claim that ⌜φ⌝ is true when these are merely hypothetically assumed) in an attempt to maximize acceptance of the compositionality intuition within the confines of classical logic. This theory is rejected principally on the grounds that it is ω-inconsistent, which, as Horsten points out, many regard as 'no more than a "sophisticated" inconsistency' (112). Horsten then considers various attempts to provide a proof-theoretic treatment of Kripke's semantic theory of truth, including the Kripke-Feferman theory (*KF*), the theory *IKF* of *KF*'s *inner* logic (i.e. not the set of those sentences contained in *KF* itself, but the set of

those sentences which *KF* claims are true), an extension *RKF* (for *restricted* KF) of the theory *PT*-mentioned above, and finally Horsten's own preferred theory, *partial* Kripke-Feferman *PKF*.

PKF is formalized within a non-classical, partial logic that does not validate the law of excluded middle; this is as one might expect, since in Kripke's theory certain sentences (such as the liar) are neither true nor have true negations. Moreover, by adopting this logic *PKF* is able to satisfy the disquotational intuition in full (138), for the deduction theorem fails, thus preventing the derivation of the inconsistent totality of all Tarski-biconditionals; and equally, it allows one to interchange the order of the truth predicate and the logical connectives quite generally, thus vindicating the compositionality intuition. Moreover, the proof-theory of *PKF* is given not with axioms, but instead using natural deduction rules; and although these rules are themselves perfectly general, they do not allow the derivation (within partial logic) of any universal laws of truth. Horsten ultimately rests his case for deflationism on this fact, suggesting that 'the absence of general laws of truth is best explained by the absence of an essence of truth' (151) beyond that which is given in our conception of it; rather, '[t]ruth is *essentially* an inferential notion' (144, emphasis original) which 'resembles the logical notions' (144), helping us to express contents we could not otherwise express, and to reason effectively and efficiently.

There is much that could be disputed in Horsten's case for deflationism and for the correctness of *PKF*: the compositionality intuition, for instance, does not strike me as unassailable; and some (e.g. logicists) might take issue with the deflationary account of logic on which Horsten's case rests. Indeed, even the assumption that Tarski's investigation of the function, rather than the essence, of truth constitutes a turn towards deflationism is not beyond question; for Frege also held that we should aim to determine the laws of truth, and he was no deflationist. But the fact that these and similar questions can – and no doubt will – be raised and debated with clarity is itself a testament to the fact that Horsten's book is a success in linking the formal and philosophical issues surrounding truth. *The Tarskian Turn* is both a valuable pedagogical contribution and an impressive scholarly achievement.