Leon Horsten. *The Tarskian Turn.* MIT Press, Cambridge, Mass., and London, 2011. $35. ISBN 978-0-262-01586-8. xii + 165 pp.

Leon Horsten has produced a valuable survey of deflationary axiomatic theories of sentential truth.

The object of study is the truth or falsity of sentences, as opposed to propositions. This raises an obvious objection: The sentence "It is snowing" isn't either simply true or simply false. It is true at some times and places and false at others. Following custom, Horsten evades this problem by restricting his attention to eternal sentences (so-called by Quine [1960, pp. 193f]), which are unaffected by contextual variation.

The axiomatic approach to theories of truth originates in Tarski's [1935, pp. 256f] observation that, in situations in which we don't have the logical resources to produce an explicit definition of truth, we will nonetheless be able to characterize the set of true sentences, simply by taking the (T)-sentences – instances of the schema "'_____' is true if and only if _____" – as axioms. Tarski made two crucial points. The first allayed fears raised by the liar paradox. Assuming that the truth predicate doesn't appear in the object language and that the object theory includes basic syntax, adding the (T)-sentences to the object theory will not render a consistent theory inconsistent. Put another way, the theory consisting of the (T)-sentences is *conservative* over the object theory: One can't prove any object-language statements with the theory that one couldn't prove without it. Second, even though the theory is materially adequate, in the sense of Convention T, the information it contains is so diffuse that it will be unsatisfactory for many purposes. Contemporary axiomatic theorists try to devise theories that are focused enough to be usefully informative and that overturn the prohibition against including the truth predicate in the object language. Allowing the truth predicate into the object language requires either restricting the range of object-language sentences that can be plugged into the (T)-schema, or else giving up

classical logic. Both possibilities are explored in *The Tarskian Turn,* although *dialetheic* theories, which allow true contradictions, are given short shrift.

Axiomatic approaches, as Horsten understands them, aim to present an effective system of axioms and/or rules of inference. He contrasts axiomatic with model-theoretic approaches, and contends [pp. 20-22] that only the former show any promise for giving us a truth theory for a natural language, because model-theoretic methods always require a metalanguage richer than the object language in which to carry out the model-theoretic construction. I don't share his confidence in axiomatics. Applications of the axiomatic method outside pure mathematics have been scarce indeed, and there's no reason to think that linguistics will be different, in this regard, from the other sciences. I suspect that the dichotomy, "Axiomatic or model-theoretic?" is overly restrictive.

What qualifies a theory as deflationary is not as clear as one would like. We are told that deflationist truth is a light and insubstantial notion [p. 3], and that it is "a concept without a nature or essence" [p. 4]. I have no idea what that means. A nearby distinction that I am able to make sense of is the division of essential from accidental attributes, but the distinction goes the opposite way from what Horsten's cryptic epigram would have led one to expect, at least according to the foremost exponent of sentential deflationism, Hartry Field. Field [1986], tells us that a substantial understanding of sentential truth will hold that a sentence has its truth conditions accidentally, as a consequence of the contingencies of speaker usage, whereas a deflationist will contend that a sentence has its truth conditions essentially. The (T)-sentences are true by stipulation, being part of the meaning of the word "true."

A more helpful suggestion is to focus on what the notion of truth useful for. For the deflationist, the most prominent uses of the notion appear, as Quine [1986, p. 11] puts it, "at

places where we are seeking generality, and seeking it along certain oblique planes that we cannot sweep out by generalizing over objects." If we have been able to affirm a great many instances of the schema "If Teacher says '_____,' then _____," and we've never encountered a counterinstance, we might hypothesize that the pattern holds generally, declaring, "Everything Teacher says is true." Without the notion of truth, we'll lack the logical resources to express the generalization. Proponents of substantial theories of sentential truth envisage a much larger role for the notion. They aim to describe, predict, and explain verbal behavior in terms of a theory of meaning within which *truth value* and *truth condition* are key theoretical concepts. (Theorists of propositional truth have a different agenda.) Horsten [p. 16] attributes to Tarski the idea that truth theorists should pay special attention to the question "What is truth useful for?" I suspect this is a misattribution.

It is by no means obvious that we can't use a single notion of truth for both purposes. There are those – McGee and McLaughlin [1995] among them – who argue that a semantic theory that grounds truth conditions in speaker usage will have to acknowledge truth-value gaps, to account for vagueness, whereas a deflationary account has no room for truth-value gaps, since "'Harry is bald' is true if and only if Harry is bald" and "'Harry is bald' is false if and only of Harry is not bald" together entail that "Harry is bald" is either true or false but not both. But this is not a majority view. Many think the two concepts of truth are coextensive.

A place where the conceptual bifurcation makes a real difference is the response to a question raised by Gödel's theorem: How do we know that the Gödel sentence – for Peano arithmetic (PA), let's say – is true? Traditionally, accepted mathematical theses fell into two categories: propositions that were regarded as obvious, and propositions that were derived from obvious propositions taken as axioms. We accept the Gödel sentence, yet it doesn't fall into

either category. An attractive answer is that the Gödel sentence is provably equivalent to CON(PA), and we know PA is consistent because we know the axioms are true, we know that the rules of inference are truth preserving, and we know that "~ 0=0" isn't true. How successful this answer is depends on our conception of truth. Although we view the Gödel sentence through the lens of a coding, so we think of it as metamathematical, in itself it's a statement of plain arithmetic, provably equivalent to the statement that a certain polynomial lacks an integer solution. We wouldn't want our acceptance of this arithmetical fact to depend on facts about the sociology of language use, such as the alleged fact that our usage gives arithmetical statements determinate truth values. People who are ambivalent about Platonism sometimes say that they accept the axioms but they don't regard them as true. If that's our attitude, we'll be forced to conclude that our acceptance of the Gödel sentence reflects a commitment to Platonism that isn't present in ordinary mathematical practice. To put the Gödel sentence on all fours with other arithmetical judgments, we'll require that the notion of truth that appears within the consistency proof be a notion that isn't encumbered with sociolinguistic or metaphysical baggage. We'll require a deflationary conception of truth. "All the theorems are true" is like "Everything the Pope says *ex cathedra* is true," something we believe, not because we've examined all the theorems or all the papal pronouncements, but because we know that they are obtained by a method – deduction in the predicate calculus in one case, divine revelation in the other – that is generally reliable.

It all fits together in a tidy package (see McGee [2006]), but Stewart Shapiro [1998] has turned the argument on its head. To complete the consistency proof, we need to know more about truth than just the (T)-sentences; Jeffrey Ketland [1999] has shown this. We need a compositional theory of truth. But the proof of the Gödel sentence shows that the compositional

theory of truth isn't conservative over PA, and if a theory isn't conservative, it isn't deflationary.

Horsten [§7.2] gets off the train at the very end. If you regard it as part of the deflationary conception that the laws of truth are true by stipulation, then it's pretty nearly obvious that they need to be conservative over your background nonsemantic theory, since you can't make substantive judgments about what the world is like true by stipulation. If, on the other hand, all deflationism requires is that the notion of truth be light and insubstantial, then it's not evident that a notion can't be light and insubstantial without being conservative.

There is an underlying dispute about the meaning of the name "Peano arithmetic." Peano himself [1889] described what we nowadays call "second-order PA." First-order PA has traditionally been understood as an axiomatic system within the language of arithmetic. This is how it was understood, for example, by Tarski, Mostowski, and Robinson [1953], and by Boolos, Burgess, and Jeffrey [2002]. An alternative view treats "PA" as contextually ambiguous. In the context of a language $\mathcal{L}$ that includes the language of arithmetic, "PA" refers to the theory we get by expanding the traditional axiom system by allowing symbols from $\mathcal{L}$ to appear within induction axioms. On the revisionist reading, to get a contextually stable proof that the Gödel sentence isn't derivable in PA, you would need to supplement the standard proof that the sentence cannot be derived arithmetically by an argument that it can't be derived by a detour through nonarithmetical language. Shapiro's argument that nonconservativeness renders a compositional understanding of truth unavailable to deflationists depends on the revised reading.

Horsten [p. 83] proposes that we resolve the disagreement by taking "PA," when we're talking about truth, unambiguously to refer to the system that we get by allowing "Tr," as well as the arithmetical symbols, to appear within induction axioms. Adding the compositional theory to PA, so understood, is arithmetically nonconservative, in that it lets us prove arithmetical

statements we couldn't prove before. This proposal makes little sense to me. Before we supply the theory of truth, there will be no connection between "Tr" and truth, so the "axioms" will be strings of symbols of an uninterpreted calculus. We can't accept them because we can't make sense of them.

In going beyond where Tarski left us, by inquiring about languages that contain their own truth predicate, we are guided by the maxim that, "Axiomatic theories of truth should be *sound*. They should prove only sentences that we instinctively and immediately accept, or, after reflective consideration, can come to see to be correct." [P 24, emphasis in original] For a book so admirable in its rigor, this is a surprisingly soft-focus aim. Moreover, a lesson we might have taken away from the liar paradox is that, when talking about truth, our instincts are unreliable. Many people find deflationism inherently unappealing. Presumably, their instincts will lead them away from the path of deflationism. How does soundness fit in with the deflationist's aim of developing a way of using "true" that is useful as a quasi-logical device?

The Tarskian Turn surveys efforts to overcome Tarski's requirement that a theory of truth for a language always has to be developed in a richer metalanguage. The survey is wide ranging, but it concentrates most of its attention on Kripke [1975], who produced a semantics with truth-value gaps in which a sentence φ is true, false, or undecided according as $\text{Tr}(\ulcorner \varphi \urcorner)$ is true, false, or undecided. Kripke's account doesn't really overturn Tarski's prohibition (as Kripke himself [1975, p. 714] notes). It is an ingenious bit of model theory, carried out within a thoroughly classical metatheory. Feferman's [1984] axiomatization of Kripke's theory is a proof theorist's delight, and Horsten is deeply immersed in proof-theoretic thinking. As much as he admires the theory, however, he cannot rest content with it, because it violates one of the demands of soundness: The theorems of a truth theory ought themselves to be, according to the theory, true.

Horsten's proposal (presented in §§ 9.5-9.6 and developed more fully in Halbach and Horsten [2006]) – although he nowhere suggests that his final word is the final word – is that our theory of truth consist of sentences that lie demonstrably inside the smallest fixed point of the Kripke construction. We get to these conclusions, not by a detour through Feferman's axioms (a course laid out carefully by Reinhardt [1986]), but by reasoning directly within the 3-valued logic. The obvious objection was raised by Feferman [1984, p. 95]: "Nothing like sustained ordinary reasoning can be carried on" within the weakened logic. Horsten responds by speculating that we could teach ourselves to reason in a new, more straitened way, and moreover, we can still help ourselves to classical reasoning when the truth predicate isn't being employed. This would be an unsurprising, if heroic, position for someone who thinks that it is pragmatically incoherent to embrace a truth theory that has theorems the theory doesn't regard as true. Once you've acknowledged truth as the norm of assertion, if you don't regard a judgment as true, you ought to refrain from asserting it. But a deflationist doesn't recognize truth as the norm of assertion. For the deflationist, truth is just a useful expository device. So why not take an easier path?

Deflationary truth was supposed to be a simple, easy notion, one we can freely employ, whatever our philosophical predilections, without metaphysics or mystery. But on the preferred account, truth is so strange and ineffable that there is nothing general we can say about it. We can't agree that the rules are truth preserving. We can't even say that every true sentence is true. This version of deflationism falls short, it seems to me, of what was advertised.

Horsten is a clear writer with a keen eye for both philosophical and mathematical insights, so, even though I was dissatisfied with the conclusion to which he brought us, I was very pleased at how much I learned along the way.

References.

Boolos, George S., John P. Burgess, and Richard C. Jeffrey. 2002. *Computability and Logic,* 4th
ed. Cambridge: Cambridge University Press.

Feferman, Solomon. 1984. "Toward Useful Type-free Theories I." *Journal of Symbolic Logic* 60:
342-345. Reprinted in Martin [1984, pp. 237-287]. Page references are to the original.

Field, Hartry. 1986. "The Deflationary Theory of Truth." In Graham MacDonald and Crispin
Wright, eds., *Fact, Science, and Morality*. Oxford: Basil Blackwell. Pp. 55-118.

Halbach, Volker, and Leon Horsten. 2006. "Axiomatizing Kripke's Theory of Truth." *Journal of
Symbolic Logic* 71: 677-712.

Ketlland, Jeffrey. 1999. "Deflationism and Tarski's Paradise." *Mind* 109: 69-104.

Kripke, Saul A. 1975. "Outline of a Theory of Truth." *Journal of Philosophy* 72: 690-716.
Reprinted in Martin [1984, pp. 53-81].

Martin, Robert L. 1984. *Recent Essays of Truth and the Liar Paradox.* Oxford: Oxford
University Press.

McGee, Vann. 2006. "In Praise of the Free Lunch." In Vincent F. Hendricks, Stig Andur
Pedersen, and Thomas Bollander, eds., *Self-Reference* (Stanford, California: CSLI, 2006),
pp. 95-120.

McGee, Vann, and Brian P. McLaughlin. 1995. "Distinctions Without a Difference." *Vagueness.
Southern Journal of Philosophy* 33 supplement: 203-51.

Peano, Giuseppe. 1889. *Arithmetices Principia, Nova Methodo Exposito.* Turin. English
translation by Jean van Heijenoort in van Heijenoort [1967, pp. 83-97].

Quine, Willard Van Orman. 1960. *Word and Object*. Cambridge, Mass.: MIT Press.

Quine, Willard Van Orman. 1986. *Philosophy of Logic,* 2nd ed. Cambridge, Mass., and London:

MIT Press.

Shapiro, Stewart. 1998. "Proof and Truth: Through Thick and Thin," *Journal of Philosophy* 95: 493-521.

Tarski, Alfred. 1935. "Der Wahrheitsbegriff in den formalisierten Sprachen." *Studia Philosophica* 1: 261-405. English translation by J. H. Woodger in Tarski [1983, pp.152-278]. Page references are to the translation.

Tarski, Alfred. 1983. *Logic, Semantics, Metamathematics*, 2nd ed. Indianapolis: Hackett.

Tarski, Alfred, Andrzej Mostowski, and Raphael M. Robinson. 1953. *Undecidable Theories.* Amsterdam: North-Holland.

Van Heijenoort, Jean. 1967. *From Frege to Gödel.* Cambridge, Mass., and London: Harvard University Press.

Department of Linguistics and Philosophy
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
vmcgee@mit.edu