TRUTH AND CIRCULAR DEFINITIONS

Francesco Orilia
Department of Philosophy, University of Cagliari (Italy)

Achille C. Varzi
Istituto per la Ricerca Scientifica e Tecnologica, Povo/Trento (Italy)

Review of Anil Gupta and Nuel Belnap, *The Revision Theory of Truth*, Cambridge, Mass., and London: MIT Press (Bradford Books), 1993, pp. xii + 299, $37.50 (cloth), ISBN 0-262-07144-4.

This original and enticing book provides a fresh, unifying perspective on many old and new logico-philosophical conundrums. Its basic thesis is that many concepts central in ordinary and philosophical discourse are inherently circular and thus cannot be fully understood as long as one remains within the confines of a standard theory of definitions. As an alternative, the authors develop a *revision theory of definitions*, which allows definitions to be circular without this giving rise to contradiction (but, at worst, to "vacuous" uses of definienda). The theory is applied with varying levels of detail to a circular analysis of concepts as diverse as truth, predication, necessity, physical object, etc. The focus is on truth, and hope is expressed that a deeper understanding of the Liar and related paradoxes has been provided: "We have tried to show that once the circularity of truth is recognized, a great deal of its behavior begins to make sense. In particular, from this viewpoint, the existence of the paradoxes seems as natural as the existence of the eclipses" (p. 142). We think that this hope is fully justified, although some problems remain that future research in this field should take into account.

The following assumptions constitute the typical background in which the truth paradoxes arise: (i) classical first-order logic, (ii) a language allowing for self-reference, and (iii) the "semantic" Tarskian schema:

(TS) $\mathsf{T}`A' \leftrightarrow A$

(where 'T' is the truth predicate, and the single quotes are a nominalization device applicable to sentences; for simplicity, we only consider homophonic versions of TS). This background can be seen as somehow part of our ordinary linguistic and conceptual background and yet, to avoid inconsistency, one or more of these assumptions must be suitably weakened. The classical, Tarskian strategy is to forbid self-reference, whereas the fixed-point approaches stemming from the work of Saul Kripke (1975) and Robert Martin and Peter Woodruff (1975) weaken the logic, allowing for truth-value gaps and giving up bivalence. By contrast, Gupta and Belnap’s basic recipe is to keep both self-reference and standard logic and to replace the unconditioned acceptance of TS with the definitional schema

(DS) $\mathsf{T}\text{‘}A\text{’} =_{\mathrm{df}} A$.

This has some independent cognitive motivations. But the move from TS to DS becomes particularly crucial insofar as the latter stands for an infinity of possibly circular definitions of the truth predicate: ‘*A*’ may stand for a sentence containing ‘T’ itself, and the symbol ‘$=_{\mathrm{df}}$’ is thus to be understood in the light of the general revision theory of definitions. This strategy—the authors argue—proves much more successful with respect to *descriptive adequacy*, i.e., the problem of avoiding inconsistency while remaining as faithful as possible to our pre-theoretical intuitions concerning truth. Indeed, descriptive adequacy appears to be the main desideratum that Gupta and Belnap are after in dealing with truth and, *mutatis mutandis*, with the other concepts they take to be circular. This makes their work relevant for all those areas of cognitive science that are interested—from various perspectives—in a formal description of our ordinary linguistic and conceptual background. In this connection, we should welcome further research on proof-theoretic systems and (possibly efficient) proof procedures adequate to the model-theoretic systems provided in the book (a class of calculi is proposed in chapter 5, “A General Theory of Definitions”). Such research is bounded by recent results by Philip Kremer (1994) and Gian Aldo Antonelli (1994a), who have shown that the two main systems discussed in the book ($\mathbf{S}^{*}$, $\mathbf{S}^{\#}$) are not axiomatizable (they have complexity Π^1_2).

Roughly, the revision theory works as follows. The crucial idea is that underlying the use of some predicates (such as truth) is not a rule of application but rather a *rule of revision*—a rule that does not fix the actual extension of the predicate, but enables us to gradually improve on some initial *hypothetical* (possibly fictitious) extension. This has no significant effect in the case of sentences that involve no circularity, for their truth value eventually stabilizes after a few revisions (thereby discharging the arbitrariness of the initial hypothesis). But the revision process becomes crucial in the presence of circularity, and can explain the pathological behavior of certain sentences. To illustrate, assume that L is a standard first-order language and M a classical model for it. Suppose we get L^{+} by enriching L with a stock of new predicates for which possibly circular definitions are provided by the set of definitions D. (For instance, we can take L^{+} to be the result of adding the truth predicate ‘T’ along with definitions patterned after DS.) In order to use M to interpret L^{+}, we start from an arbitrary hypothesis concerning the interpretation of the new predicates and set off a revision process in an attempt to interpret each of them as demanded by the corresponding definiens. A hypothesis provides a classical interpretation for each definiendum in D, i.e., a classical truth value **t** or **f**, given an *n*-adic definiendum and any *n*-tuple drawn from the domain of M. In other words, given M, a hypothesis h gives rise to a classical model $M+h$ for L^+. If we assume an arbitrary hypothesis h_0 and model $M+h_0$ as a starting point, the revision process generates a revision sequence of models $M+h_0$, $M+h_1$, $M+h_2$, …, by means of a revision rule δ that takes as input a hypothesis h_n and gives as output a new hypothesis h_{n+1}. At each successor level n+1, δ assigns to each definiendum the set of n-tuples satisfying the corresponding definiens in the previous model $M+h_n$. At each limit stage α, if a definite verdict on the interpretation of a definiendum G^n has been reached (in the sense that, from a certain point onward, each new hypothesis always assigns the same truth value to the pair constituted by G^n and a given set of n-tuples of D), this verdict is preserved in the new hypothesis $\delta(h_\alpha)$. These cases give rise to sentences that are called *stably true* or *stably false* (relative to M) as the case may be. Whenever no such verdict has been reached (*unstable* sentences), different options present themselves, giving rise to alternative revision theories. Some such alternative options have been explored in the previous literature on revision-theoretic approaches to the theory of truth (by Belnap (1982), Gupta (1982, 1988/1989), and Hans Herzberger (1982)). This book reconsiders them from the wider perspective of the revision theory of definitions, and compares them with some novel treatments proposed here for the first time. (Perhaps some weakness in the exposition may be noted here. The authors take good care in explaining the intuitive rationale behind the "preliminary" systems $\mathbf{S}_n$, but the intuitions behind the other systems are somewhat left for the reader to sort out.)

In dealing specifically with truth, three different model-theoretic systems ($\mathbf{T}^*$, $\mathbf{T}^\#$, $\mathbf{T}^c$) are proposed, and, with descriptive adequacy in mind, it is shown how their relying on a classical two-valued semantics allows them to capture intuitively valid informal arguments that cannot be formalized by the competing approaches. (The motto is, "The addition of a truth predicate to a language does not disturb the logical structure of the language in any way", p. 142.) Roughly, the three systems differ as follows. System $\mathbf{T}^*$ is based on the idea that unstable sentences get an arbitrary truth-value at limit stages; $\mathbf{T}^\#$ lifts this arbitrariness for unstable sentences whose truth-values at worst fluctuate only for a finite segment after limit ordinals (nearly stable sentences); finally, $\mathbf{T}^c$ relies on the principle that the extension of the truth predicate should always be a maximally consistent set of sentences. The three systems are not equally successful with respect to descriptive adequacy, but Gupta and Belnap do not make any definite commitment. This raises the question of which of these systems should be regarded as *the* theory referred to in the title of the book. System $\mathbf{T}^\#$ fares better than the others, but does not guarantee that it can be freely used without giving rise to ω-inconsistency. In view of a result of Vann McGee (1985), this problem cannot be removed without giving up to some extent "semantic principles" such as

(T~) $\quad \mathrm{T}`{\sim}A' \leftrightarrow {\sim}\mathrm{T}`A'$, that contribute to the success of $\mathbf{T}^{\#}$ in meeting descriptive adequacy (p. 225). We thus face a difficult dilemma. The authors argue at some length that ω-inconsistency is not as “bad” as it might seem at first sight. But we think future research should try to further refine the notion of descriptive adequacy in order to deal with such dilemmas.

This applies to more specific results as well. For instance, the authors point out that there are intuitively valid arguments that are not captured by any of the systems proposed (cf. example 6C.10, p. 228). The problem with these arguments is that they would require a prima facie correct appeal to TS. It would thus be worth characterizing interesting classes of cases for which this principle can be safely upheld in an attempt to agree as much as possible with pre-theoretical intuitions. Regarding example 6C.10, it is shown that it could be successfully tackled by a system in which, at limit stages, only “fully varied revision sequences” (p. 168) are taken into account, but unfortunately no such system is actually constructed. To develop it and experiment with it should contribute to further enhancing the general approach proposed by Gupta and Belnap. (We are told there is a paper on this by André Chapuis forthcoming in the *Journal of Philosophical Logic*).

There are also some controversial principles with respect to which the proposed systems do not remain neutral. As an example, if ‘ℓ’ is a simple Liar sentence, then the disjunction

(1) $\quad \mathsf{T}‘\ell’ \quad {\sim}\mathsf{T}‘\ell’$

comes out stably true even if both disjuncts are paradoxical. This is so for reasons vaguely reminiscent of the motivations that led supervaluational semanticists to accept the Law of Excluded Middle while rejecting Bivalence (pp. 261–263; compare Kit Fine’s supervaluational treatment of vagueness in (1975)). As Steve Yablo (1985) already pointed out in connection with Gupta’s and Herzberger’s early formulations, this reflects one chief hidden assumption of the revision approach, viz., that the hypotheses over which the revision procedure randomizes include one that is *correct*. If such hypotheses are all possible classical interpretations of the truth predicate, then (1) follows for supervaluational reasons. But this seems far from being uncontroversial.

A related example is

(2) $\quad \mathsf{T}‘\ell’ \quad \ell$

which is validated by all systems discussed in the book. This cannot be claimed to be an undesirable outcome on purely intuitive grounds. But we suspect that some discussion of this and similar results could further clarify the basic mechanisms operating in the revision method. (Such sentences are validated also by modified versions of Gupta and Belnap’s theories, as long as ℓ and ${\sim}\mathsf{T}‘\ell’$ are made to coincide. For instance, Aladdin Yaq¯;u b’s system (1993), which is designed precisely to overcome problems arising with similar artifacts, treats (2) as valid.)

This also relates to what Haim Gaifman (1992) calls the “black hole” problem, viz., the fact that no information concerning the truth-value of a pathological sentence can be stated directly. For instance, the revision theory does not distinguish between:

(3) ~T(3)
(4) ~T(3),

in spite of the obvious difference (the former, but not the latter, is self-referential). The authors argue that “any assertion that ‘the Liar is untrue’, even when made with the full consciousness of the Liar’s paradoxicality, invites the response that the Liar must then be true, since it asserts its own untruth. The circle of semantical reflection is not naturally broken at any point” (p. 255, fn. 5). It would, however, be interesting to see this point further developed. For instance, this is a point where the basic assumption of treating truth as a predicate of sentences (p. 12)—i.e., sentence types, as opposed to sentence tokens (which is what is peculiar about (3) and (4))—deserves careful examination.

These examples do not, in our view, weaken the interest and richness of the material presented in the book. However, they are indicative of the difficult issues hidden behind the authors’ choice to emphasize the role of descriptive adequacy, particularly in view of the claim that “we should abandon the primacy of formal correctedness: A definition should be evaluated only by how well it captures the material aspects of a notion” (p. 277).

More examples and open problems are discussed in the last chapter of the book (“Truth and Other Circular Concepts”), which is worth reading before going through the technical details of the preceding chapters. This final part also emphasizes the generality of the revision-theoretic method, showing that it can be uniformly applied to a variety of other topics in addition to truth. The authors give some illustrations in connection with other semantic concepts, such as reference and satisfaction, as well as with set-theoretic, property-theoretic, modal, and doxastic notions. We believe there is room for much development here, and some results are already appearing in the literature. For instance, Antonelli (1994b) has used revision rules to construct models of set theory with non-well-founded sets. Other applications, we believe, are forthcoming, and will show all the potentials of the revision theory apart from whatever specific misgivings one may have. This impressive and technically accomplished book must be considered a must for any reader with serious interests in the fundamental questions of logic and semantics and their cognitive underpinnings.

(The book is well edited, and there are no substantial typos. Two minor exceptions: on p. 65, line 7, the supremum sign should be replaced by the infimum sign; on p. 66, second line of 2C.6, ‘po’ should read ‘ccpo’.)

References

Antonelli, Gian Aldo (1994a), 'The Complexity of Revision', *Notre Dame Journal of Formal Logic* **35**, 67–72.

Antonelli, Gian Aldo (1994b), 'Non-Well-Founded Sets via Revision Rules', *Journal of Philosophical Logic* **23**, 633–679.

Belnap, Nuel D. (1982), 'Gupta's Rule of Revision Theory of Truth', *Journal of Philosophical Logic* **11**, 103–116.

Fine, Kit (1975), 'Vagueness, Truth, and Logic', *Synthese* **30**, pp. 265–300.

Gaifman, Haim (1992), 'Pointers to Truth', *Journal of Philosophy* **89**, 223–261.

Gupta, Anil (1982), 'Truth and Paradox', *Journal of Philosophical Logic* **11**, 1–60.

Gupta, Anil (1989), 'Remarks on Definitions and the Concept of Truth', *Proceedings of the Aristotelian Society 1988/89* **89**, 227–246.

Herzberger, Hans G. (1982), 'Notes on Naive Semantics', *Journal of Philosophical Logic* **11**, 61–102.

Kremer, Philip (1993), 'The Gupta-Belnap Systems $\mathbf{S}^{\#}$ and $\mathbf{S}^{*}$ are not Axiomatisable', *Notre Dame Journal of Formal Logic* **34**, 583–596.

Kripke, Saul (1975), 'Outline of a Theory of Truth', *Journal of Philosophy* **72**, 690–716.

Martin, Robert L., and Peter R. Woodruff (1975), 'On Representing 'True-in-L' in L', *Philosophia* **5**, 217–221.

McGee, Vann (1985), 'How Truth-Like Can a Predicate Be? A Negative Result', *Journal of Philosophical Logic* **14**, 399–410.

Yablo, Steve (1985), 'Truth and Reflection', *Journal of Philosophical Logic* **14**, 297–349.

Yaq¯;u b, Aladdin M. (1993),*The Liar Speaks the Truth*, Oxford: Oxford University Press.