THE TRUTH TELLER PARADOX

Chris MORTENSEN
Graham PRIEST

1. Consider the statements

(α) This very sentence is true

(β) This very sentence is false

The second of these, (β), is of course the liar paradox, about which much has been written. Much less has been written about (α), though many have noted that it is curious. One aspect of that curiousness is that there seems to be nothing to choose between the hypotheses that it is true, and that is is false. Indeed, it is hard to see how there could even be anything to choose between the hypotheses. More particularly, both hypotheses seem to be consistent: from neither hypothesis does there appear to be deducible a contradiction. And if that is true, it would seem further to follow that there is no *a priori* proof of the truth of (α), and no *a priori* proof of its falsity. (Contrast this situation with that of (β), where there are, *prima facie* at least, *a priori* proofs of both the truth and the falsity of (β).) We will return to the question of the absence of *a priori* proofs of the truth and of the falsity of (α) below, in section 6.

2. However, it has gone unnoticed that (α) has a specific problem of its own. The purpose of this note is to explain the problem and to consider in a little more detail some aspects of it, particularly some possible solutions. The problem, which we might call «the truth teller paradox» is as follows. In view of the fact that truth and falsity are both consistently assignable to (α), and so there are no *a priori* proofs of its truth and its falsity, it is tempting to suppose that it is neither true nor false. However, this runs into the following problem. Suppose that (α) is neither true nor false. Then certainly it is not true. But since it says of itself that it is true, it is false. This contradicts the

supposition that it is neither true nor false. Hence it is either true or false.

To put the problem in a nutshell, the absence of any possible proof of the truth of (α) would be a reason for declaring it not to be true. Similarly, there is a reason to declare it not to be false. In any case, there are apparently no proofs of its truth or of its falsity. But there is a proof that it is either true or false.

3. There are several options for formalising, in the interests of clarity, the argument that (α) is either true or false. Not all of these options are equivalent. We choose one which seems to us to be closest to the intuitions expressed in the above informal argument. (The notation is the obvious one).[1]

Let	$\ulcorner\alpha\urcorner = \ulcorner T\ulcorner\alpha\urcorner\urcorner$	(1)
Now	$(\sim T\ulcorner\alpha\urcorner \,\&\, \sim F\ulcorner\alpha\urcorner) \rightarrow \sim T\ulcorner\alpha\urcorner$	(2)
But	$\ulcorner\alpha\urcorner = \ulcorner T\ulcorner\alpha\urcorner\urcorner \rightarrow (\sim T\ulcorner\alpha\urcorner \rightarrow F\ulcorner\alpha\urcorner)$	(3)
So	$(\sim T\ulcorner\alpha\urcorner \,\&\, \sim F\ulcorner\alpha\urcorner) \rightarrow F\ulcorner\alpha\urcorner$	(4) By (1), (2), (3)
	$\rightarrow (T\ulcorner\alpha\urcorner \vee F\ulcorner\alpha\urcorner)$	(5) From (4)
But	$T\ulcorner\alpha\urcorner \vee F\ulcorner\alpha\urcorner \vee (\sim T\ulcorner\alpha\urcorner \,\&\, \sim F\ulcorner\alpha\urcorner)$	(6)
Hence	$T\ulcorner\alpha\urcorner \vee F\ulcorner\alpha\urcorner$	(7) By (5) and (6)

4. We do not know how to solve the truth teller paradox, though some solutions are more attractive than others (see section 7 below). However, it will serve to sharpen the paradox, we think, if we chart briefly where the possibilities for solution lie. The first possibility is to fault one of the lines of the argument, and in this section we consider this. Line (1) just formalises the English sentence (α). Lines (2), (4), (5) and (7) depend on standard truth functional properties of conjunction and disjunction, (specifically $(A \,\&\, B) \rightarrow A$, $A \rightarrow (A \vee B)$ and $(A \rightarrow B) \rightarrow ((A \vee B) \rightarrow B)$), the transitivity of entailment, and modus ponens for $\rightarrow$. Most of these have been questioned.[2] In our view they all hold, and it seems fair to say that most philosophers would be in agreement.

That leaves lines (3) and (6). Line (3) is an expression of the idea that if (α) fails to be true, then, in the light of the fact that it says of itself that it is true, it is false. That seems correct to us, and pretty much on all fours with the more usual argument, in the liar paradox,

that if (β) is true, it is false. Evidently, however, someone might refuse to make that move, and simply hold to the proposition that (α) fails to be true. Of course, that is not quite so simple a matter, because there are various arguments one can imagine to back up (3). For instance, from $\sim T^{\ulcorner}\alpha^{\urcorner}$, using $^{\ulcorner}\alpha^{\urcorner} = T^{\ulcorner}\alpha^{\urcorner}$, we have, by substitution of identicals, $\sim\alpha$; then, by the T-scheme (i.e. $\alpha \leftrightarrow T^{\ulcorner}\alpha^{\urcorner}$), $T^{\ulcorner}\sim\alpha^{\urcorner}$; so that $F^{\ulcorner}\alpha^{\urcorner}$ (by definition of «false», say). Holding that (α) merely fails to be true requires that one find fault with some move in this latter argument. That is not so easy, and doubtless various philosophers could be found with opposing intuitions on just where to halt the argument. For example, one position holds that the T-scheme holds only for «grounded» sentences, and that neither (α) nor its denial are grounded. (3) Because this kind of position is frequently taken with respect to logical paradoxes like (β), its strengths and weaknesses are well known. (4) We will therefore not pursue the matter except to say that we think that the weaknesses outweigh the strengths, as one of us has, in effect, argued elsewhere. (5)

That leaves (6): either (α) is true, or it is false, or it is neither. This is an expression of one of the fundamental ideas behind the paradox, that if one wants to avoid holding that (α) is true and one wants to avoid holding that it is false, then one should say that it is neither. We think it would be a mistake to assume the Law of Excluded Middle in the form: (α) is true or (α) is false; but of course (6) does not depend on that, and seems to be on much more solid ground. It is possible to deduce (6) from an instance of Excluded Middle in the form $A \vee \sim A$:

$$(T^{\ulcorner}\alpha^{\urcorner} \vee F^{\ulcorner}\alpha^{\urcorner}) \vee \sim(T^{\ulcorner}\alpha^{\urcorner} \vee F^{\ulcorner}\alpha^{\urcorner}), \text{ so } T^{\ulcorner}\alpha^{\urcorner} \vee F^{\ulcorner}\alpha^{\urcorner} \vee (\sim T^{\ulcorner}\alpha^{\urcorner} \,\&\, \sim F^{\ulcorner}\alpha^{\urcorner})$$

by De Morgan's Law. But it is preferable to avoid that if possible, if only because it is a tricky business working out precisely what Excluded Middle comes to in a situation where we are taking seriously the possibility that some sentence might be neither true nor false.

5. If all the steps of the argument are accepted there appear to be only three further possibilities. The first is a paraconsistent or inconsistent one, namely to accept both the argument to the effect that (α) is either true or false *and also* the arguments to the effect that (α) has no truth value. Thus $\sim(\sim T^{\ulcorner}\alpha^{\urcorner} \,\&\, \sim F^{\ulcorner}\alpha^{\urcorner}) \,\&\, (\sim T^{\ulcorner}\alpha^{\urcorner} \,\&\, \sim F^{\ulcorner}\alpha^{\urcorner})$ would be a true contradiction. While it is arguable that this position

can be sustained without collapse into triviality,(6) it would tempt few.

6. The second possibility is to accept that either (α) is true or (α) is false, but deny that there is any paradox by holding that one of those two alternatives really is true, but (at least at present) we do not know which. After all, in section 1 all we said was that the hypotheses that (α) is true and that it is false *seem* to be consistent; and from that it does not follow that there *is* no proof of the truth of (α), nor that there is no proof of its falsity. Now clearly the truth of one of the disjuncts is not going to be determined empirically, so if a truth value for (α) is to be found it will be by the production of an *a priori* proof.

Indeed, that is not so unthinkable. Although there appears to be nothing to choose between (α) being true and its being false, appearances can be deceptive. As an analogy, consider the sentence normally expressed as «This sentence is provable in Peano arithmetic». It might well be thought that this would have exactly the same ontological status as (α). Yet it is provable in Peano arithmetic.(7) It might also be thought that some standard proof of this fact could be modified by substituting «is true» for «is provable» to produce an argument for the truth of (α). Indeed, that modification can be made.

Let «δ» be the claim «If this sentence is true, α».

	i.e. $\ulcorner\delta\urcorner = \ulcorner T\ulcorner\delta\urcorner \rightarrow \alpha\urcorner$
So certainly	$\delta \rightarrow (T\ulcorner\delta\urcorner \rightarrow \alpha)$
Hence	$T\ulcorner\delta\urcorner \rightarrow (T\ulcorner T\ulcorner\delta\urcorner\urcorner \rightarrow T\ulcorner\alpha\urcorner)$ by the T scheme and the distribution of T over $\rightarrow$.
But by the T scheme	$T\ulcorner\delta\urcorner \leftrightarrow T\ulcorner T\ulcorner\delta\urcorner\urcorner$; so $T\ulcorner\delta\urcorner \rightarrow (T\ulcorner\delta\urcorner \rightarrow T\ulcorner\alpha\urcorner)$
Hence by contraction	$T\ulcorner\delta\urcorner \rightarrow T\ulcorner\alpha\urcorner$
and by the T scheme	$T\ulcorner\delta\urcorner \rightarrow \alpha$ (*)
and again	$T\ulcorner T\ulcorner\delta\urcorner \rightarrow \alpha\urcorner$
which by definition of α gives	$T\ulcorner\delta\urcorner$
whence by line (*)	α

Unfortunately this is unsatisfactory as an *a priori* proof of (α). It is just a dressed-up version of Curry's Paradox and could easily be

modified to produce a proof of $\sim\alpha$. It cannot be carried out in Peano arithmetic, because the truth predicate for Peano arithmetic fails to be arithmetically representable (at least, it fails iff Peano arithmetic is consistent!). In natural language, anyone who wants to maintain that the truth predicate is representable([8]) will presumably make similar moves against this proof of (α) that he or she would make against Curry's Paradox.([9])

But still, even though this proof fails, there may be satisfactory proofs of either $T\ulcorner\alpha\urcorner$ or $F\ulcorner\alpha\urcorner$. However, perhaps there are no proofs of either $T\ulcorner\alpha\urcorner$ or $F\ulcorner\alpha\urcorner$. In this case we would be forced into the position that something true is in principle unprovable. This is certainly ruled out from an Intuitionist viewpoint, though not, in principle, from a classical one. However, it does raise the question of in what *exactly* the truth of one of these alternatives consists.

7. The final possibility for solving the truth teller paradox is the one which seems to us to be the most attractive, if only because the difficulties sketched in the foregoing solutions appear considerable. This solution is to hold that while (α) fails to be true and fails to be false, nevertheless it is either true or false. The possibility of this line goes back to Aristotle.([10]) Adopting the position requires that the standard truth conditions for disjunction be rejected (even though we set out to use standard disjunction), since we have all of: (1) (α) is true $\vee$ (α) is false, (2) (α) is not true, and (3) (α) is not false. Intensional disjunction has been studied, for example by Anderson and Belnap.([11]) There, however, the intensional disjunction is not the kind for which $A \vee B$ might hold while both A and B fail, but the kind for which the truth of one of the disjuncts is not sufficient for the truth of the disjunction.

A theory is said to be *prime* iff whenever $A \vee B$ is in the theory either A is in the theory or B is. Identifying the world with its true theory, we can then describe the present position by saying that it holds that *the world is non-prime*. The failure of primeness is not mysterious. For example, let PA be Peano arithmetic formulated with a base of classical logic, and let G be its Gödel sentence. Now certainly $\vdash_{PA} G \vee \sim G$. But, by Gödel's first Incompleteness Theorem, if PA is consistent then neither $\vdash_{PA} G$ nor $\vdash_{PA} \sim G$. That is, (the set of theorems of) Peano arithmetic is non-prime if it is consistent.

A typically Platonist attitude might interpret Gödel's theorems as applying to consistent recursively enumerable arithmetics only, while holding that arithmetical reality is complete. If arithmetical reality were consistent and complete, then it would be prime (as a simple argument establishes). But one man's modus ponens is another man's modus tollens: a consistent and non-prime reality fails to be complete. The failure of the completeness of the world is also a doctrine which has been taken seriously.

We cannot pretend that the thesis that the world is non-prime is not paradoxical; but it is, we believe, no more paradoxical in principle than the paraconsistent solution to the liar, namely that (β) is both true and false([12]) (and paradoxes not infrequently beget paradoxical solutions). In favour of both solutions, it can be said that our concepts of negation, falsity and disjunction *escape our control*. This is especially so in logico-mathematical contexts such as the present ones where, as with Intuitionism, truth and falsity seem to coincide with provable truth and falsity. Consider, for example, the liar (β). We might feel that we can *resolve* to make «false» exclude «true», but this is an illusion. There is no guarantee that we can keep control of the resolution once we allow «false» and «true» to have additional logical properties. If we allow them to have enough properties, particularly the natural property of the representability of the truth predicate in the object language, then we will *discover* something about the concepts, namely that we cannot keep to the resolution of the exclusiveness of «true» and «false». Furthermore, such sentiments ought not to be so strange to someone with Intuitionist leanings. It only needs the view that in limited contexts, say logico-mathematical contexts, something is true (false) iff provably true (false), for it to seem not so impossible that the truth value of a sentence might be overdetermined. (Of course, it is at the very least arguable that the assignment of both truth and falsity to a sentence does not lead to logical collapse.)([13]) Similarly, then, with the truth teller, (α). We cannot guarantee control of the desirable primeness feature of disjunction if we want disjunction to have additional, natural properties.

Supposing that primeness fails does not mean that the truth conditions for «∨» go completely haywire. Nothing has been said to deny that the truth of disjuncts is *sufficient* for the truth of a disjunction. As to *necessity* here, we can usefully invoke the idea of

the local consistency and completeness of a theory. Let L be a language, and Th a theory in that language. Then Th is *locally consistent (complete) relative to a sublanguage L' of L* iff the restriction of Th to L' is consistent (complete). It follows from the previously cited fact (that consistency and completeness implies primeness), that local consistency and completeness implies local primeness. Normal, well-behaved situations will be consistent and complete, we can suppose. Hence, a locally well-behaved theory of the world will be one in which disjunction is locally entirely classical. It follows that we may be confident that it is only when things get strange that the truth conditions for «∨» behave unexpectedly. One might, in addition, have a theory about just when strangeness like inconsistency, incompleteness and non-primeness can occur (say, in logico-mathematical contexts, though we do not wish to commit ourselves to such a restrictive view).

8. In sum, the truth teller paradox is interesting in its own right. In addition, at least some solutions raise intriguing possibilities.

Australian National University Chris MORTENSEN
University of Western Australia Graham PRIEST

FOOTNOTES

(1) We let «α» be an abbreviation for the sentence (α), not its name. Quasi quotes will be used as name forming functors.

(2) On (A & B)→A, see Storrs MCCALL «Connexive Implication», *Journal of Symbolic Logic*, 31 (1966), 415-433. On A → (A ∨ B) and transitivity see William PARRY «Ein Axiomsystem für eine neue Art von Implikation (Analysische Implikation)», *Ergebnisse eines mathematischen Kolloquiums*, 4 (1933), 5-6. On transitivity see Peter GEACH *Logic Matters*, Blackwell, 1972, Chs. 4,7. On modus ponens, see Errol MARTIN and Robert K. MEYER, *S for Syllogism*, forthcoming; or Robert K. Meyer, Richard Routley and J. Michael Dunn, «Curry's Paradox», *Analysis*, 39 (1979), 124-8.

(3) See Saul KRIPKE, «Outline of a Theory of Truth», *Journal of Philosophy*, 12 (1975), 690-716.

(4) See e.g. Susan HAACK, *Philosophy of Logic*, Cambridge U.P., 1978, 145 ff.

(5) See Graham PRIEST, «The Logic of Paradox». *The Journal of Philosophical Logic*, 43, (1978).

(6) PRIEST, *ibid*; or Nicholas Rescher and Robert Brandom, *The Logic of Inconsistency*, Blackwell, 1980.

([7]) See e.g. George BOOLOS and Richard JEFFREY, *Computability and Logic*, Cambridge U.P., 1974, Ch. 16, 188-9.

([8]) See PRIEST, *ibid.*

([9]) For example, by giving up contraction $(A \rightarrow (A \rightarrow B)) \rightarrow (A \rightarrow B)$, or modus ponens in the form $((A \rightarrow B) \& A) \rightarrow B$. See MEYER, ROUTLEY and DUNN, *ibid.*

([10]) *De Interpretatione*, Ch. 9. Formally it might be handled by supervaluation techniques. See e.g. BAS C. VAN FRAASSEN «Presupposition, Implication and Self-Reference», *Journal of Philosophy*, 65 (1968), 136-52.

([11]) For example, Alan Ross ANDERSON and Nuel BELNAP, *Entailment*, Princeton, 1975, 176-7.

([12]) See PRIEST, *ibid.*

([13]) See PRIEST, *ibid.*