Truth and Inconsistent Concepts

John Barker
University of Illinois Springfield

Are the semantic paradoxes best regarded as formal puzzles that can be safely delegated to mathematical logicians, or do they hold broader philosophical lessons? In this paper, I want to suggest a philosophical interpretation of the liar paradox which has, I believe, nontrivial philosophical consequences. Like most approaches to the liar, this one has deep roots, having been first suggested by Tarski (1935) and later refined by Chihara (1979)\(^1\). I offered a further elaboration of the idea in Barker (1999), and here I would like to develop these ideas a bit further.

The term 'liar paradox' refers to the fact that the ordinary disquotational properties of truth -- the properties that allow semantic ascent and descent -- are formally inconsistent, at least on the most straightforward way of formally expressing those properties and given standard assumptions about the background logic. The best-known formulation of those disquotational properties is Tarski's convention (T):

(T) \( 'A' \) is true if and only if \( A \)

We now consider a sentence such as

(1) Sentence (1) is not true.

As long as the schematic letter \( A \) in (T) has unlimited scope, we can derive the following instance:

(2) 'Sentence (1) is not true' is true if and only if sentence (1) is not true.

Then, noting that the sentence quoted in (2) is none other than sentence (1) itself, we derive the consequence

(3) Sentence (1) is true if and only if sentence (1) is not true.

And this conclusion, (3), is classically inconsistent: it is an instance of \( P \leftrightarrow \neg P \).

The liar paradox should concern all of us, because it represents a gap in our understanding of truth, and because truth is a central notion in philosophy, mathematical logic and computer science. Tarski's (1935) work on truth is what finally put mathematical logic on a firm foundation, and led to the amazing explosion of work in that field. Tarski's work in turn inspired Davidson (1967), whose influential work gives truth a central place in semantic theory. And computer science, of course, is based on mathematical logic; the theory of computability itself is essentially just the theory of truth for a certain fragment of the language of arithmetic\(^2\). (For more on the relation between logic and computability see Hintikka's (2011) contribution to this Newsletter.) If truth plays such an important role in all three fields, then it behooves us to get to the bottom of the paradoxes.
There is now a truly vast body of literature on the liar, and the argument (1-3) above is far from the last word on the subject. Having said that, the liar paradox is remarkably resilient. Accounts of the liar can be divided into two camps: descriptive and revisionary. For a revisionary account, the goal is to produce a predicate with disquotational properties of some sort, which can serve the purposes that we expect a truth predicate to serve, while not necessarily being wholly faithful to our naive truth concept. This approach has much to recommend it. But in this paper, I will be focusing on descriptive accounts. If the ordinary notion of truth needs to be replaced by a revised notion, I want to know what it is about the ordinary notion that forces us to replace it. If the ordinary notion is defective in some sense, I want to know what it means to say it is defective. And if, on the other hand, we can produce an account of truth that avoids contradiction and is wholly faithful to the ordinary concept, then there is no need to go revisionary.

Descriptive accounts, in turn, can be divided into the following categories, depending on what they hope to achieve.

- **Block the contradiction.** Descriptive accounts in this category proceed from the assumption that there is a subtle but diagnosable flaw in the reasoning that leads to contradictions such as (3). Indeed, it's not hard to convince oneself that there must be such a flaw: if an argument has a contradictory conclusion, there must be something wrong with its premises or its inferences.

- **Embrace the contradiction.** On this approach, there's nothing wrong with the reasoning leading up to the conclusion (3). That conclusion simply expresses the fact that the liar sentence (1) is both true and not true. This approach, known as dialetheism, has never been the majority view, but lately it has received a surprising amount of attention.

- **Acknowledge the contradiction.** On this approach, convention (T) is part of the meaning of 'true', and so the contradiction (3) is in some sense a consequence of the concept of truth. This differs from "embracing" the contradiction in that the contradiction (3), while viewed as a commitment of ordinary speakers, is not actually asserted. This will be the approach taken here.

Revisionary accounts also try to block the contradiction; and if the contradiction can be effectively blocked, then doing so is the preferred approach, I would think. But blocking the contradiction turns out to be hard, especially (I will argue) in the context of a descriptive account. In the next section, I will explain some of the reasons why this is the case. If blocking the contradiction is as hard as I think it is, we should at least entertain the alternatives, provided the alternatives are intelligible at all. In the remainder of this paper, I will try to explain what it means to acknowledge the contradiction, and why it makes sense to do so.

### 1. Why the Liar Is Hard

Any account of the liar, whether descriptive or revisionary, has to operate within the following constraint:

**Constraint 1.** The truth predicate, as explained by the theory at hand, must have the expected disquotational properties.
And this by itself is not easy to achieve: we saw earlier that a natural formulation of the "expected disquotational properties" led directly to a contradiction. Having said that, there is some wiggle room when it comes to "expected disquotational properties", and we also have some leeway in our choice of background logic. In fact, there are theories of truth that have some claim to satisfying Constraint 1.

Let's consider a couple of examples: not the highest-tech examples, to be sure, but sufficient for our purposes. First, Tarski's original proposal was to simply restrict convention (T) so that the substituted sentence \( A \) is forbidden from containing the truth predicate. Then the substitution of sentence (1) for \( A \) is prohibited, and the contradictory conclusion (3) cannot be derived. But this restriction on (T) is quite severe, limiting what we can do with the resulting truth predicate even in a revisionary account. For a descriptive account, Tarski's restriction is simply a non-starter, since natural language clearly places no such limit on what can substitute for \( A \) in (T). (And it should be noted that Tarski himself viewed this approach as revisionary, not descriptive.)

Another approach to revising (T), which results in a less severe restriction, starts from the idea that not all sentences are true or false. In particular, some sentences represent truth value gaps, with the liar sentence (1) a very plausible candidate for such treatment. If gaps are admitted, then we can maintain an equivalence between the sentences \( A \) and ' \( A \) is true' for all \( A \) in our language. In particular, when \( A \) is gappy, so is ' \( A \) is true'. The first mathematically rigorous treatment along these lines is due to Kripke (1975), who developed a family of formal languages containing their own gappy truth predicates, each obeying a suitable version of (T). Sentences like (1) can then be proved to be gappy in Kripke's system.

The main weakness of Kripke's approach is that the languages in question need to be developed in a richer metalanguage. Some of the key notions of the account, while expressible in the metalanguage, are not expressible in the object language. In particular, the notion of a gappy sentence, which is obviously crucial to the account, has no object language expression. The reason is simple and instructive. On the one hand, in Kripke's construction, there is an object language predicate \( Tr \), and it can be shown that \( Tr \) is a truth predicate in the sense that (a) an object language sentence is true if and only if it belongs to \( Tr \)'s extension, and (b) an object language sentence is false if and only if it belongs to \( Tr \)'s anti-extension. (Predicates in Kripke's system have extensions and anti-extensions. A predicate \( P \) is true of those objects in its extension, false of those in its anti-extension, and neither true nor false of anything else.) Now suppose the object language had a gappiness predicate as well. That is, suppose there were a predicate \( G \) whose extension included all and only the gappy sentences. We could then construct a sentence that says "I am either not true or gappy": i.e., a sentence \( S \) that is equivalent to \( \neg Tr(\neg S) \lor G(\neg S) \). \( S \), like any sentence, is either true, false or gappy. But if \( S \) is true, then both \( \neg Tr(\neg S) \) and \( G(\neg S) \) are not true, and thus neither is \( S \). If \( S \) is false, then \( \neg Tr(\neg S) \) is true, and thus so is \( S \). And if \( S \) is gappy, then \( G(\neg S) \) is true, and hence so is \( S \). So \( S \) is neither true, false, nor gappy, which is impossible. This contradiction (in the metatheory) proves that no such predicate as \( G \) exists.

Kripke described this phenomenon as the "ghost of the Tarskian hierarchy", because despite his efforts to create a self-contained object language, he found it necessary to ascend to a richer metalanguage, just as Tarski did. The problem is also called the strengthened liar problem, because the sentence \( S \) is a "strengthened," i.e., harder to deal with, version of the liar sentence, and also as the revenge problem, since the moment we account for one manifestation of the liar problem, a new manifestation appears to take revenge on us. The key feature of the revenge
problem is that in addressing the liar, we develop a certain set of conceptual tools (in this case, the notion of a truth value gap); those tools are then turned against us, i.e., they are used to construct a new liar sentence (in this case, $S$) which our original account is unable to handle.

Whatever we call it, the revenge problem shows that even though Kripke was able to construct an internally consistent way of satisfying truth's expected disquotational properties, he did so at the expense of placing a tacit restriction on the sorts of sentences that the resulting truth predicate applies to. Specifically, he constructed a truth predicate for a language in which the metalanguage notion of gappiness is inexpressible. The construction used to create the strengthened liar sentence $S$ is rather general, and the prima facie lesson of the revenge problem is that an account of truth can't be given for the language in which the account is formulated.

If this is so -- and so far it has been suggested but not proved -- then it is moderately bad news for revisionary accounts and extremely bad news for descriptive accounts. From a revisionary perspective, the revenge problem simply means that in constructing a predicate with the desired disquotational properties, we will have to be content with a predicate that applies only to a certain fragment of the language we speak. Some sentences in our language may be assertible, and we may even be committed to asserting them, but we can't use our (revisionary) truth predicate to describe them as true: they simply fall outside that predicate's scope. This might be a limitation we can live with. But from a descriptive perspective, it is puzzling. The ordinary concept of truth applies, or at least it certainly seems to apply, to all sentences of our language, not just to some formally tractable fragment of our language. That is, descriptive accounts have to live with the following additional constraint.

**Constraint 2.** A descriptive account of truth must describe a truth predicate for an entire natural language, not just a fragment of a natural language.

So suppose we have an account of truth, and suppose it uses some notion, like gappiness, that doesn't occur in the sentences to which the truth predicate, as described by our theory, applies. In what language is this account stated? The natural obvious answer is: it is stated in a natural language, e.g., English. But then what we have produced is an account of truth for a proper fragment of English, not for all of English, in violation of Constraint 2.

For this reason, it has often been suggested that when we formulate an account of truth, we sometimes do so not in an ordinary language like English, but in a richer language, call it English+. English+ is English supplemented with technical terms, like 'gappy', that are simply not expressible in ordinary English. And the resulting account is a theory of true sentences of English, not of English+. Such a move faces some challenges, however.

First of all, if one holds that English+ is needed to formulate a theory of truth for English, then it is hard to resist the thought that a still-further enhanced language, English++, could be used to formulate a theory of truth for English+. The process can clearly be iterated, leading to a sequence of ever-richer extensions of English, each providing the means to express a theory of truth for the next language down in the hierarchy. We can even say exactly how this works: English+ comes from English by adding a predicate meaning 'gappy sentence of English'; English++ comes form English+ by adding a gappy-in-English+ predicate; and in general, for each language $L$ in the hierarchy, the next language $L+$ is obtained from $L$ by adding a predicate for the gappy sentences of $L$.

However, once we have all this on the table, a question very naturally arises: what language are we speaking when we describe the whole hierarchy of languages? Our description
of the hierarchy included the fact that English+ has a predicate for gappiness in English; but 'gappy in English' is not expressible in English, so our account must not have been stated in English. Parallel reasoning shows that our account cannot have been stated in any language in the hierarchy. We must have been speaking some super-language English* that sits at the top of the entire hierarchy. And then we're right back where we started from, since clearly we need a theory of truth for English* as well.

Maybe a better approach is to just drop talk of the hierarchy of languages, or at most to understand it as a form of Wittgensteinian gesturing rather than rigorous theorizing. But there is another problem. Let's just focus on the languages English and English+, where again English+ is the result of adding a predicate to English that means 'gappy sentence of English'. English+ is, again, the metalanguage in which we diagnose the liar paradox as it arises in English. This approach assumes that the truth predicate of English applies only to sentences of English: English has a predicate meaning 'true sentence of English', but does not have a predicate meaning 'true sentence of English+'. If it did, then that predicate could be used to construct a gappiness predicate in English. Specifically, we could define 'gappy sentence of English' in English as follows:

\[
A \text{ is a gappy sentence of English if and only if the sentence 'A is gappy' is a true sentence of English+}.
\]

And since English does not have a gappy-in-English predicate -- the entire approach depends on this -- it doesn't have a true-in-English+ predicate either. More generally, if English had a true-in-English+ predicate, then English+ would be translatable into English, which is impossible if English+ is essentially richer than English. So any theory of truth that, by its own lights, can only be stated in an essentially richer extension English+ of English must also maintain that (ordinary) English lacks a truth predicate for this extended language.

All of which sounds fine, until one realizes that the truth predicate of English (or of any other natural language, I would think) is not language-specific. The truth predicate of English purports to apply to propositions, regardless of whether they are expressible in English or not. This should actually be obvious. Suppose we discovered an alien civilization, and suppose we had good reason to suspect that the language they speak is not fully translatable into English. Even if we assume this is the case, it does not follow that the non-translatable sentences are never used to say anything true. On the contrary, it would be reasonable to assume that some of the extra sentences are true. But then there are true sentences that can't be expressed in English. Or suppose there is an omniscient God. Then it follows that all of God's beliefs are true; but it surely does not follow that all of God's beliefs are expressible in English.

So the ordinary truth predicate applies, or purports to apply, to sentences of any language, and this fact forms another constraint on descriptive accounts:

**Constraint 3.** The truth predicate, as described by the account, must apply to sentences of arbitrary languages (or to arbitrary propositions).

But this constraint is incompatible with the richer-metalanguage approach. To see this, suppose 'gappy sentence of English' really is expressible only in some richer language English+. This means that some people -- some philosophers who specialize in the liar, for example -- actually speak English+. Let Bob be such a speaker. That is, let 'Bob' be a term of ordinary English that
denotes one such speaker. ('Bob' could abbreviate a definite description, and there are plenty of those in ordinary English.) Then we can say, in ordinary English, for any phoneme or letter sequence $A$,

(4) The sentence '$A$ is gappy' is true in Bob's idiolect.

If 'true' behaves the way it intuitively seems to, as described in Constraint 3, then (4) is true in English if and only if $A$ is gappy in English. So English has a gappiness predicate after all, which directly contradicts the account we have been considering.

For these reasons, I think an account of truth that requires a move to a richer metalanguage is unpromising as a descriptive account, however much value it might have as a revisionary account. So what are the prospects for a descriptive account that does not require a richer metalanguage? A complete answer would require a careful review of the myriad accounts in the literature, a monumental undertaking. But let me offer a few observations.

First, because the problem with expressing gappiness is a formal problem, it is relatively insensitive to how the gaps are interpreted. Because of this, numerous otherwise attractive proposals run into essentially the same revenge problem. Here are some examples.

Truth is a feature of propositions, and the liar sentence fails to express a proposition. This is an attractive way of dealing with liar sentences, until one realizes that failing to express a proposition is just a way of being gappy, and that the usual problems with gappiness apply. The strengthened liar sentence, in this case, is

(5) Sentence (5) does not express a true proposition.

Does sentence (5) express a proposition? First, suppose not. Then a fortiori, (5) does not express a true proposition. In reaching this conclusion, we used the very words of (5): we wound up assertively uttering (5) itself. And in the same breath, we said that our very utterance failed to say anything. And our account committed us to all this. This seems to be an untenable situation, so maybe we should reconsider whether (5) expresses a proposition. But if (5) does express a proposition, then that proposition must be true, false or gappy (if propositions can be gappy), any of which leads to trouble. Here's another example:

There are two kinds of negation that occur in natural language, wide-scope and narrow-scope (or external and internal). In the liar sentence (1), the negation used is narrow-scope. When we step back and observe that (1) is not true, our 'not' is wide-scope.

Well and good, but the natural and obvious response is to simply construct a liar sentence using wide-scope or external negation:

(6) Sentence (6) is not\textsuperscript{wide} true.

Then, in commenting that (6) is gappy and thus not true, we are assertively uttering the same words as (6) in the very same sense that was originally intended.
A perennially\textsuperscript{6} popular response is to regard truth ascriptions as ambiguous or otherwise context-sensitive, and to diagnose the liar on that basis. The intuition behind this response is as follows. We would like to say that (1) is gappy, and being gappy is a way of not being true. So we reach a conclusion that we express as follows:

(7) Sentence (1) is not true.

Formally, sentence (7) is the same as the liar sentence (1), and so in assertively uttering (7), we are labeling the words of our very utterance as not true. Intuitively, though, there seems to be an important difference between the utterances (1) and (7). In (7), we are stepping back and evaluating (1) in a way that we weren't doing with (1) itself. This has led some philosophers to suggest that (1) and (7) actually say different things.

The tools to formally express this idea go back to the Tarskian hierarchy of languages and, before that, the Russellian hierarchy of types. Using Burge's (1979) account as an example, suppose we explain differences like that between (1) and (7) in terms of differences in the content of 'true' on different occasions. That is, suppose we treat 'true' as indexical. Let's use numerical subscripts to mark the different extensions of 'true': true\textsubscript{1}, true\textsubscript{2}, .... Then sentence (1), fully subscripted, is rendered as follows:

(1') (1') is not true\textsubscript{1}.

On an account like Burge's, (1') is indeed not true: i.e., it is not true\textsubscript{1}. We express this in the same words as (1'):

(7') (1') is not true\textsubscript{1}.

But in assertively uttering (7'), don't we commit ourselves to the truth of (7')? Indeed we do, but not to the truth\textsubscript{1} of (7'). From (7'), what we are entitled to conclude is

(8) (7') (and thus (1')) is true\textsubscript{2}.

And there is no conflict between (7') and (8). Problem solved! A bit more formally, what we have done is modify the disquotational properties of truth somewhat. We have, for any given sentence $A$ and index $i$,

(T\textsubscript{i,1}) If 'A' is true\textsubscript{i}, then $A$

And we have a weak converse: for any $A$, there exists an index $i$ such that

(T\textsubscript{i,2}) If $A$, then 'A' is true\textsubscript{i},

This modified disquotational principle is perfectly consistent, and on the face of it, it leaves us with a perfectly serviceable disquotational device.

One question that can be raised about such a proposal is whether there is any evidence, aside from the paradoxes themselves, that the natural language word 'true' really works this way.
I do think this is a worry, but there is another, potentially more damaging problem. Consider the following sentence, sometimes called the "super-liar":

(S) Sentence (S) is not true, for any $i$

Using ($T_i^1$), it is easily seen that (S) is not true, for any $i$. That is, (S) is not true at all: there is no context in which it is correct to say that (S) is true. And yet our conclusion here -- sentence (S) is not true, for any $i$ -- is stated in the very words of (S), so there had better be some sense in which (S) is true. Thus, we have what seems to be a violation of ($T_i^2$).

The standard response is that (S) is simply ill-formed: it relies on binding the subscript $i$ with a quantifier, which is not permitted. This response is correct as far as it goes, but it misses the fact that (S) is a well-formed sentence of the metalanguage in which the account is presented. Or at least, something with the same gist as (S) can be expressed in the metalanguage. After all, the account at issue makes explicit generalizations about the hierarchy of truth predicates, for example the claims ($T_i^1$) and ($T_i^2$). Such claims presuppose some mechanism for generalizing across indices, and once that mechanism is in place, we can use it to construct sentences like (S). Indeed, (S) and ($T_i^1$) are entirely parallel: each is (or can be written as) a schema with a schematic letter $i$, understood as holding for all indices $i$. If you can say ($T_i^1$) in the metalanguage, you can say (S) too.

But we plainly can't say (S) in the object language, so we're back to the problem of the essentially richer metalanguage. Notice also that the problem of (S) is a classic example of the revenge problem: the machinery of the account -- in this case, the ability to generalize across indices -- is used to construct a new liar sentence that the account can't handle.

In summary, we have found some substantial obstacles to a satisfactory descriptive account of truth, at least if that account is to satisfy the three constraints mentioned above; and those constraints are certainly well-motivated. What are we to make of this?

2. The Inconsistency Theory

One possible response to these considerations is to simply reject one or more of Constraints 1-3. However, there are different things that it can mean to reject a constraint. It might be that at least one of the constraints is simply factually wrong: the natural language truth predicate doesn't work like that, even though it seems to. Alternatively, we could argue that while the constraints are in fact part of the notion of truth, there is no property that satisfies these constraints, and hence, no such property as truth. My proposal will be somewhat along the latter lines; but let's first consider the former proposal.

One could certainly reject one or more of the constraints of the last section as factually incorrect, but such a move seems to me to be very costly. Suppose, for example, that we reject Constraint 1, that truth has the expected disquotational properties. For example, suppose we maintain that in some special cases, assertively uttering a sentence does not carry with it a commitment to that sentence's truth. This would free us up to assert, for example, that

(9) (1) is not true
without worrying that this will commit us to the truth of (9) (and hence, of (1)): the above sentence may simply be an exception to the usual disquotational rule.

But one seldom finds such proposals in the literature, and I think the reason is clear: the disquotational principles seem to be part of the meaning of 'true'. One might even say they seem analytic. And this consideration seems to have a lot of pull, even with philosophers who don't believe in analyticity. Finding a sentence that turns out to be an exception to the disquotational rules would be like finding a father who is not a parent. The disquotational rules seem to me to be so much a part of our notion of truth that rejecting them would be tantamount to declaring that notion empty.

Likewise, one could question whether a descriptive theory needs to apply to the language it's stated in. That is, one could reject Constraints 2 and 3. But this would be tantamount to claiming that the ordinary notion of truth applies only to a proper fragment of the language we speak, or at least a proper fragment of a language we could (and some of us do) speak, and it seems clear that truth, in the ordinary sense, has no such limitation.

Yet another possibility is to simply accept the existence of truth value gluts: of sentences that are both true and not true. This at least has the virtue of simplicity. Convention (T) can be taken at face value, and there's no need for complicated machinery or richer metalanguages. As for the costs of this approach, many would consider its commitment to true contradictions to be a cost in itself.

But suppose we could get the explanatory benefits of dialetheism without being saddled with true contradictions. That is, suppose there were a way to maintain that (T), or something like it, really is part of the concept of truth, without actually claiming that liar sentences are both true and untrue. Such an account might be very attractive.

Along these lines, let's start with a thought experiment. Imagine a language where nothing serves as a device of disquotation. The speakers get together and decide to remedy the situation as follows. First, a string of symbols is chosen that does not currently have a meaning in the language. For definiteness, let's say the string in question is 'true'. Next, the following schema is posited, with the intent of imparting a meaning to this new word:

(T) \( 'A' \) is true if and only if \( A \).

It is understood that \( A \) should range over all declarative sentences of the language, or of any future extension of the language. And that's it: positing (T) is all our speakers do to impart any meaning or use to 'true'. The word 'true' goes on to have a well-entrenched use in their language, long before anyone realizes that contradictions can be derived from (T).

There are a number of observations we can make about this thought experiment. First, it is coherent: we can easily imagine a group of speakers doing exactly what I have described. We can certainly debate what meaning, if any, the word 'true' has in their language, but it seems clear that a group of speakers could put forward (T) with the intention of giving a meaning to the new word 'true'.

Second, we can easily imagine that the positing of (T) leads to 'true' having a well-defined use in the speakers' language. We simply have to imagine that 'is true' is treated as a predicate and that the application of (T) as an inference rule becomes widespread. We might even imagine that once the use of 'true' becomes well-entrenched, the explicit positing of (T) fades from memory -- but that's getting a bit ahead of the story.
Third, in saying that the speakers establish a use for 'true', we should understand 'use' in a normative sense, as governing the correct use of 'true', and not just as summarizing speakers' actual utterances or dispositions to make utterances. This is crucial if we want to say that (T) has a special status in the language, and isn't just a pattern that the speakers' behavior happens to conform to. It is also the sort of thing we should say in general: the notion of use that is relevant to questions of meaning, I claim, is the normative sense. In any case, I think it's clear from the thought experiment that (T) is put forward as a norm, and adopted as a norm by the speakers.

Fourth, I claim that the positing and subsequent uptake of (T) confers a meaning on 'true', in some sense of 'meaning'. Here we have to be careful, because the word 'meaning' itself has several different meanings, and 'true' (in this example) may not have a meaning in every sense. It's not obvious, for example, that 'true' has a well-defined intension. What I mean is that 'true' in the imagined case is not simply nonsense; it plays a well-defined role in the language.

Fifth, and finally, there is nothing in this thought experiment that forces us into dialetheism in any obvious way, even if we accept the foregoing observations. We've simply told a story about a language community adopting a certain convention involving a certain word; doing so shouldn't saddle us with any metaphysical view about things being both so and not so. To put it a bit differently: there's nothing contradictory in our thought experiment in any obvious way, so we can accept the scenario as possible without thereby becoming committed to true contradictions. Of course, the speakers themselves are, in some sense, committed to contradictions, specifically to the contradictory consequences of (T); but that's a separate matter. There's a big difference between contradicting yourself and observing that someone else has contradicted herself.

It should come as no surprise that I think the above thought experiment bears some resemblance to the actual case of the word 'true' in English. However, there is an important difference between the two cases. Namely, no natural language ever got its truth predicate from an explicit positing of anything like (T). We shouldn't read too much into this difference, however. In the thought experiment, the initial stipulation of (T) plays an important role, but an even more important role is played by the speakers' incorporation of (T) into their language use. Eventually, the fact that (T) was stipulated could fade from memory, and any interesting feature of the word 'true' would depend on its ongoing use. In which case the question arises: what interesting feature does 'true' have in these speakers' language?

The best answer I know is that the speakers have a language-generated commitment to (T), which was initially established by the act of positing (T) and then sustained by the speakers' ongoing use of 'true'. I think this accurately describes the language of the thought experiment, and I suggest that (aside from the business about positing) it describes natural languages as well. In the case of natural language, (T) is not an explicit posit, but it is a convention of language, accepted tacitly like all such conventions.

So this is the inconsistency theory of truth as I propose it. In natural languages, there is a language-generated commitment to the schema (T) or something very much like it. Using (T), we can reason our way to a contradiction. This gives rise to the liar paradox, and it explains why the liar is so puzzling: we don't know how to block the reasoning that generates the contradiction, because the reasoning is licensed by our language and our concepts themselves.

As evidence for the inconsistency theory, I would make the following points. First, the considerations of the previous section should make an inconsistency theory worth considering. Second, the inconsistency theory is simple: no elaborate gyrations are required to avoid paradox, either in our semantic theory or in the conceptual schemes we attribute to ordinary speakers. And
third, the inconsistency theory does justice to the sheer intuitiveness of (T). My native speaker intuitions tell me that (T) is analytic; and the inconsistency theory supports this intuition. Indeed, if one were to accept the inconsistency theory, it would be very natural to define a sentence to be analytic in a given language if that language generates a commitment to that sentence.

The inconsistency theory shares these virtues with dialetheism, which is unsurprising given the similarity of the two views. But (as I will argue at greater length in the next section) the inconsistency doesn't actually have any contradictory consequences. For those philosophers (like me) who find true contradictions a bit hard to swallow, this should be an advantage.

3. Refinements, Objections and Ramifications

Is the inconsistency theory any different from dialetheism, though? We need to know, that is, whether the inconsistency theory implies that the liar is both true and not true, or more generally, whether it implies both $P$ and not $P$ for any $P$. Equivalently, we need to know whether the inconsistency theory is an inconsistent theory.

One might argue that the present account makes logically inconsistent claims about obligations. On our account, we have a language-generated commitment to (T). This means that at least in some circumstances, we have an obligation to assert (T)'s instances, as well as the logical consequences of (T)'s instances. Thus, we have an obligation to assert that the liar sentence (1) is true, and we also have an obligation to assert that (1) is not true. Now if the logic of negation also generates a prohibition on asserting both $A$ and not $A$ -- as I think it does -- then we have a case of conflicting obligations. And, it can be objected, this latter claim is itself inconsistent.

What this objection gets right is that the inconsistency theory regards the language-generated commitment to (T) as a kind of obligation, and not (or not just) as a kind of permission. It's not that we are licensed to infer $A$ from 'A is true' and vice versa, but need not make this inference if we don't feel like it: if we assert $A$, we are thereby committed to 'A is true', and are therefore obligated to assert 'A is true', at least in those circumstances where we need to express a stance on the matter at all. Moreover, the obligations in question are unconditional: they have no hidden escape clauses and can't be overridden like Ross-style prima facie obligations.

The only proviso attached to the commitment to (T) is that it is conditional upon speaking English, and specifically on using 'true' with its standard meaning. We can always use 'true' in a nonstandard way, or even refrain from using it altogether, working within a 'true'-free fragment of English. The point of the present account is that if we choose to go on using 'true' with its ordinary meaning, then we are thereby committed to (T).

So is it inconsistent to say that a given act is both obligatory and prohibited? For whatever reason, this matter seems to be controversial; but I think there are many cases where conflicting obligations of just this sort clearly do occur. Case 1. A legislature can create a law mandating a given act $A$, or it can create a law prohibiting $A$. What if it (unknowingly) did both at once? Then the act $A$ would be both obligatory and prohibited under the law. Case 2. People can enter into contracts and thereby acquire obligations. People can also enter into contracts with multiple third parties. What if someone is obligated to do $A$ under one contract, but prohibited from doing $A$ under a different contract? Case 3. Games are (typically) based on rules, and a poorly crafted set of rules can make inconsistent demands on the players. As a simple example,
imagine a variation on chess -- call it chess* -- with the following additional rule: if the side to move has a pawn that threatens the other side's queen, then the pawn must capture the queen. The trouble with this rule is that in some cases, the capture in question is illegal, as it would leave the king exposed. But it is certainly possible for people to adopt the rules of chess* anyway, presumably unaware of the conflict. In that case, there will eventually be a case in which a move is both required and prohibited.

Each of the examples just cited involves a kind of social convention, and so we have reasons for thinking that conventions can sometimes make inconsistent demands on their parties. If language is conventional in the same sense, then there should be a possibility of inconsistent rules or conventions of language as well. (The biggest difference is that in language, the terms of the convention are not given explicitly. But why should that matter?) In all cases of inconsistent rules, since one cannot actually both perform a given act and not perform it, some departure from the existing rules must take place. The "best" such departure is, arguably, to revise the rules and make them consistent. But this isn't always feasible (and pragmatically may not always be desirable), so the alternative is to simply muddle through and do whatever seems the most sensible. Either way, the response is inherently improvisational. It may be worth noting here that when presented with a case of the liar, most people do in fact just muddle through as best they can, in a way that seems to me to be improvisational rather than rule based. In any case, I don't think there is any inconsistency in the claim that a given system of obligations includes conflicts.

Another possible source of inconsistency for the present account is as follows. If the inconsistency theory is right, then speakers of English are committed to (a) the truth of the liar sentence (1), and (b) the non-truth of (1). That theory, moreover, is stated in English. Doesn't that mean the theory itself is committed to both the truth and the non-truth of (1)?

No, it doesn't. To see this, consider that while I did use English to state the inconsistency theory, in principle I needn't have. I could have stated the account in some other language -- say, a consistent fragment of English. In that case, anyone who wants to assert the theory without also being committed to inconsistent sets of sentences need only confine herself to some consistent language in which the theory is statable. If this is possible -- if there is a consistent language in which the inconsistency theory can be stated -- then the act of asserting the theory need not be accompanied by any commitment to a contradiction, and therefore the theory itself does not imply any contradiction.

To put this point a bit differently, if the inconsistency theory is true, then we as speakers of English are committed to both the truth and the non-truth of (1). But this doesn't imply that the theory itself is committed to the truth and non-truth of (1). The theory takes no stand on that issue. As speakers of English, we may feel compelled to take some stand on the issue, and indeed, as speakers of English we may be obligated to take conflicting stands on the issue. But it doesn't follow that the inconsistency theory itself takes any particular stand.

This all assumes that there is a consistent language -- a consistent fragment of English, or otherwise -- in which the inconsistency theory can be stated. If there isn't, then the inconsistency theory arguably becomes self-defeating or degenerates into dialetheism. This will be a problem if, and as far as I can see only if, the inconsistency theory requires the (ordinary) notion of truth for its formulation. Does it?

An old argument against inconsistency theories, due to Herzberger (1967), is as follows. Consider the claim that two sentences A and ~A are analytic. This will be the case if A and ~A are both logical consequences of some self-contradictory analytic sentence B, where B might be a contradictory instance of (T), for example. The classic definition of analyticity is as follows: a
sentence is analytic if it is true by virtue of its meaning. In particular, an analytic sentence is true. But then we have that both $A$ and $\sim A$ are true. Furthermore, we presumably have that $\sim A$ is true if and only if $A$ is not true. In that case, we have shown that $A$ is both true and not true. Thus, the claim that a sentence $B$ is both analytic and contradictory is itself a contradictory claim. Finally, if the inconsistency theory is the claim that the instances of (T) are analytic, then by Herzberger's argument, the inconsistency theory is inconsistent.

In response, I never actually claimed that (T) is analytic, and more importantly, if I were to do so I certainly would not not use the above definition of analyticity. In fact, I do think that 'analytic' is an apt term for the special status of (T), but only if analyticity is understood in terms of language-generated commitments and not in terms of truth by virtue of meaning. As an aside, there's nothing sacred about the "true by virtue of meaning" definition of analyticity, which historically is only one of many.

A similar objection, also made by Herzberger, runs as follows. The inconsistency theory is a theory about the meaning of the word 'true'. Meaning is best understood in terms of truth conditions, or more generally of application conditions. But what, then, are the application conditions of the ordinary word 'true'? That is, what is the extension of 'true'? The answer cannot be: the unique extension that satisfies (T), since there is no such extension. There seems to be no way to explain (T)'s special status in truth-conditional or application-conditional terms.

I think it's pretty clear, then, that the inconsistency theory, while a theory of meaning, cannot be understood as a theory of anything resembling truth conditions. And this raises the broader question of how the present account fits into the more general study of language.

Truth conditional semantics, of course, represents just one approach to meaning. A theory based on inferential role semantics (as per Brandom (1994)) might accommodate the present account easily. Roughly speaking, inferential role semantics explains the meaning of an expression in terms of the inferences it participates in with respect to other expressions. The cases where inferential role semantics is most convincing are those of logical operators, with the associated inference rules providing the inferential role. The inconsistency theory of truth fits easily within this framework, provided the inferences can be inconsistent -- and why can't they be? Moreover, the truth predicate strikes many as a logical operator, with the inferences from $A$ to 'A is true' and vice versa appearing to many (myself included) as logical inferences, suggesting that the truth predicate ought to be a good candidate for inferentialist treatment.

Of course, not everyone is an inferentialist, and indeed some sort of truth-conditional approach may be the most popular take on meaning. To those who are sympathetic to truth conditions (myself included!), I make the following suggestion. Facts about truth conditions must somehow supervene on facts about the use of language. How this takes place is not well understood, but may be thought of, roughly speaking, as involving a "fit" between the semantic facts and the use facts. Moreover, I suggest that these use facts should be understood as including normative facts, including facts about commitments to inferences. (These facts, in turn, must somehow supervene on still more basic facts, in a way that is not well understood but which might also be described as "fit".) Now in the case of an inconsistent predicate such as 'true', the expected semantic fact -- in this case, a fact about the extension of the predicate -- is missing, because no possible extension of the predicate fits the use facts sufficiently. (Any such extension would have to obey (T), and none does.) We might describe this as a breakdown in the language mechanisms that normally produce referential facts. I would suggest that there are other, similar breakdowns in language, such as (some cases of) empty names. Be that as it may, while there
isn't much useful we can say about the ordinary predicate 'true' at the semantic level, we can still say something useful at the use level, namely that there is a commitment to (T).

This is what I think we should say about inconsistent predicates in general, though there is a snag when the predicate in question is 'true'. Namely, on the account just sketched, the semantic facts include facts about reference and truth conditions. But if the use of 'true' is governed by an inconsistent rule and lacks a proper extension, what sense does it make to talk about truth conditions at all? This is indeed a concern, but it assumes that the notion of truth that we use when talking about truth conditions is the same as the ordinary notion of truth that this paper is about. It need not be. In particular, I have been stressing all along the possibility of a revisionary notion of truth, and it may well be that one of the things we need a revisionary notion for is semantic theory. The feasibility of this project -- i.e., of finding a paradox-free notion of truth that can be used in a semantic theory -- is obviously an important question. Fortunately, there is a great deal of contemporary research devoted to this problem.

Let me end by describing two competing views of language. On one view, a language provides a mapping from sentences to propositions. Speakers can then use this mapping to commit themselves to various propositions by assertively uttering the corresponding sentences. Language determines what we can say, and only then do speakers decide what gets said. The language itself is transparent, in that it doesn't impose any commitments or convey any information. In short, a speaker can opt into a language game without taking on any substantive commitments. I think this is a rather widespread and commonsensical view, but it is incompatible with the inconsistency theory. On that theory, speaking a natural language commits one to (T) and to (T)'s consequences, which are substantive. The medium and the message are less separate than the commonsense view suggests. This actually strikes me as a welcome conclusion -- (T) is just one of many ways, I suspect, that the language we speak incorporates assumptions about the world we speak of -- but it may also be one reason why the inconsistency theory is not more popular.

References


---

1 Similar ideas were also expressed by Carnap (1937); see especially sec. 60. While the first systematic development of the idea seems to be that of Chihara, the general notion of an inconsistency theory of truth was well known after Tarski's work, and there was sporadic discussion in the literature; see especially Herzberger (1967).

2 Specifically, a set or relation is recursively enumerable iff it can be defined in the fragment of the language of arithmetic whose logical operators are &, ∨, ∃x, and ∀x<y. A set or relation is recursive (i.e., computable) iff it and its complement are recursively enumerable.

3 See, e.g., Priest (2006).

4 It was suggested, for example, by Kripke in (1975), and later defended in detail by Soames in (1999).

5 See Soames (1999) for an account along these lines.

6 See Russell (1908), Parsons (1974), and Burge (1979), among others.