Between Logic and Intuition

Essays in Honor of Charles Parsons

Edited by

GILA SHER
University of California, San Diego

RICHARD TIESZEN
San Jose State University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK http://www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA http://www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain

First published 2000

Printed in the United States of America

Typeface Times Roman 10/12 pt. *System* LaTeX 2_ε [TB]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication Data is available.

ISBN 0 521 65076 3 hardback

Contents

Paradox Revisited I: Truth

HILARY PUTNAM

In this pair of essays, I revisit the logical paradoxes. In the present essay I discuss the most famous of the so-called semantical paradoxes, the paradox of the Liar, the sentence that says of itself that it is not true, and in the essay that follows (Paradox Revisited II) I shall consider whether we should really accept a view once expressed by Gödel, the view that the paradoxes of set theory are ones that we can see through, can definitely and satisfactorily resolve, even if (as he conceded) the same cannot be said for the semantical paradoxes.

The Liar Paradox

The best presentation I know of the Liar Paradox is Charles Parsons', and in the end the view I shall defend is, I believe, an elaboration of his. In "The Liar Paradox,"[1] a paper I have thought about for almost twenty years, the paradox is stated in different ways. One of these ways is in terms of three alternatives: either a sentence expresses a true proposition, or it expresses a false proposition, or it does not express a proposition at all. A second way mentioned in that paper is the one I followed in my presentation of the Liar paradox in *Realism with a Human Face*,[2] in which talk of propositions is avoided, and I mostly employ that way here in order to facilitate comparison with Tarski's work.

It is an empirical fact that the one and only sentence numbered (I) on page 11 of my *Realism with a Human Face* is the following:

(I) The sentence (I) is false.

Is the sentence numbered (I) (on page 11 of my *Realism with a Human Face*) true? Tarski famously used "Snow is white" as his example of a typical sentence, and his "Convention T" requires that a satisfactory treatment of truth must enable us to show that

"Snow is white" is true if and only if snow is white.

If we suppose that sentence (I) has a truth value at all, it follows by Convention T that

(i) "The sentence (I) is false" is true if and only if the sentence (I) is false.

But, as just mentioned, sentence (I) = "The sentence (I) is false," and hence

(ii) Sentence (I) is true if and only if sentence (I) is false

which is a contradiction!

So far, we do not have an actual inconsistency. We assumed that sentence (I) has a truth value, and that assumption has now been refuted. We cannot consistently assert either that (I) is true or that (I) is false. But now we come to the "strong Liar." The form I considered in *Realism with a Human Face* (p. 12) is:

(II) The sentence (II) is either false or lacks a truth-value.

Sentence (II) is paradoxical because, if we try to avoid the previous argument by denying that (II) has a truth value, that is, by asserting

(II) lacks a truth value,

then it obviously follows that

(II) is either false or lacks a truth value,

and sentence (II) is one that we discover ourselves to have just asserted! So, we must agree that (II) is true, which means that we have contradicted ourselves.

Tarski showed us how to avoid such paradoxes by relativizing the predicate "is true" to whichever language we are speaking of, and by introducing a hierarchy of languages. If I say of a sentence in a language L that it is true or false, my assertion belongs to a language of a higher level – a meta-language. No language is allowed to contain its own truth predicate. The closest I can come to such sentences as (I) or (II) is to form a sentence (III) with a relativized truth predicate:

(III) The sentence (III) is not true-in-L,

but this sentence does not belong to L itself, only to meta-L. Since it does not belong to L, it is true that it is not true-in-L. And since this is exactly what it says in meta-L, it *is* true in meta-L. Sentence (III) is not even well formed in the "object language" L, and is true in the meta-language, meta-L, and this dissolves the paradox.

In *Realism with a Human Face*, I asked "if Tarski [had] succeeded, or if he ha[d] only pushed the antinomy out of the formal language and into the informal language which he himself employs when he explains the significance of his formal work." If each language has its own truth predicate, and the notion "true-in-L," where L is a language, is itself expressible in meta-L, but not in L, all of the semantical paradoxes can be avoided, then I agreed. "But in what language is Tarski himself supposed to be saying all this?" I asked. (p. 13)

"Tarski's theory introduces a "hierarchy of languages," I continued.

> There is the "object language" . . . . there is the meta-language, the meta-meta-language, and so on. For every finite number *n*, there is a meta-language of level *n*. Using the so-called transfinite numbers, one can even extend the hierarchy into the transfinite – there are meta-languages of higher and higher *infinite* orders. The paradoxical aspect of Tarski's theory, indeed of any hierarchical theory, is that one has to stand outside the whole hierarchy even to formulate the statement that the hierarchy exists. But what is this "outside place" – "informal language" – supposed to be? It cannot be "ordinary language," because ordinary language, according to Tarski, is semantically closed and hence inconsistent. But neither can it be a regimented language, for no regimented language can make semantic generalizations about itself or about languages on a higher level than itself. (pp. 13–14)

I also considered Parsons' way out; as I explained it then, this way involves the claim that the informal discourse in which we say such things as "every language has a meta-language, and the truth predicate for the language belongs to the meta-language and not to the language itself" is not part of any language but a kind of speech that is *sui generis* (call it, "systematic ambiguity"). And I found difficulty in seeing what this comes to. After all, one can formally escape the paradox by insisting that all languages properly so-called are to be written with ink other than red, I pointed out, and red ink reserved for discourse that generalizes about "languages properly so-called." Since generalizations about "all languages" would not include the Red Ink Language in which they are written (the Red Ink Language is *sui generis*), we cannot derive the Liar paradox. But is this not just a formalistic trick? How, I asked, does Parsons' "systematic ambiguity" differ from Red Ink Language? In this essay, I hope to answer my own objection, and thereby to deepen our understanding of what systematic ambiguity is and why it is necessary. In the essay that follows, I will, among other things, argue that something like systematic ambiguity is inevitable in set theory as well.

"Black Hole" Sentences?

When I wrote the title paper of *Realism with a Human Face*, I regarded Tarski's hierarchical solution as just a technical solution, a way of constructing restricted languages in which no paradox arises. It seemed to me that, as a *general* solution, the hierarchical solution can only be "shown but not said"; it is literally inexpressible. But I proposed no solution of my own. In an earlier paper, a memorial lecture for James Thompson, I did propose a solution, but I was dissatisfied and did not publish that lecture. In that unpublished lecture, I set up a language that is not hierarchical, and in which the truth predicate can be applied to any and all of the sentences of the language. Semantical paradoxes were avoided by assuming the Convention T only for a subset of the sentences of the language, the "Tarskian" sentences. I did not define the set of Tarskian sentences once and for all; instead there were axioms enabling us to prove that certain sentences (sentences that are sure to be paradox free and that are likely to be needed) are all Tarskian. The idea was that, just as we add stronger axioms of set existence to set theory when we discover that we need them, we could add axioms specifying that additional sentences are Tarskian as these become necessary. With respect to the obviously paradoxical sentences such as the Liar, the position I recommended was a sort of logical quietism; that is, "Don't say anything (semantical) about them at all!" (This idea was, perhaps, an anticipation of Haim Gaifman's idea that there are "black hole" sentences, that is, sentences that are paradoxical and, moreover, such that the application of any semantical predicate to one of them simply generates a further paradox.[3]) But this seemed to me desperately unsatisfactory, for if we are content not to say anything at all about the paradoxical sentences, why do we not just stick to Tarski's solution? The problem with that solution, after all, arises only if we try to state it as a *general* solution. One could just as well be a quietist about the principle *underlying* the Tarskian route to avoidance of the paradoxes in particular cases as about the Black Hole sentences. Both forms of "quietism" are *so* unsatisfactory that I want to make another attempt to see if we can find something more satisfactory to say about the paradoxes. But I must warn you in advance that what I will end up with will not be a "solution" to the paradoxes, in the sense of a point of view that simply makes all appearance of paradox go away. Indeed, I still agree with the main moral of "Realism with a Human Face," which is that such a solution does not seem to be possible.

Reconsideration of Parsons' Solution

As I indicated, I now accept Parsons' solution. As I explained a moment ago, my objection to that solution was that it rests on the notion of systematic ambiguity, but it wasn't clear to me why systematic ambiguity wasn't just another language, a language that was simply *stipulated* to be outside the hierarchy, and thus available to serve as a kind of Archimedean point. Another problem (one that I did not mention in "Realism with a Human Face") was that I had difficulty in understanding the following: when Parsons applies his solution to natural language, he asserts that

> However vaguely defined the schemes of interpretation of the ordinary (and also not so ordinary) use of language may be, they arrange themselves naturally into a hierarchy, though clearly not a linearly ordered one. A scheme of interpretation that is "more comprehensive" than another or involves "reflection" on another will involve either a

larger universe of discourse, or assignments of extensions or intensions to a broader body of discourse, or commitments as to the translations of more possible utterances. A less comprehensive interpretation can be appealed to in a discourse using the more comprehensive interpretation as a metadiscourse.

To many the hierarchical approach to the semantical paradoxes has seemed implausible in application to natural languages because there seemed to be no division of a natural language into a hierarchy of "languages" such that the higher ones contain the "semantics" of the lower ones. Indeed there is no such neat division of any language as a whole. What the objection fails to appreciate is just how far the variation in the truth-conditions of sentences of a natural language with the occasion of utterance can go, and in particular how this can arise for expressions that are crucially relevant to the semantic paradoxes: perhaps not "true," but at all events quantifiers, "say," "mean," and other expressions that involve indirect speech. (p. 250)

The fact is that I found it difficult to understand what Parsons meant by his claim that talk of "meaning" and "saying" and "expressing" in natural language (if not the use of the predicate "true" itself) presupposes interpretations that can be arranged into hierarchies. But what I want to do now is to apply Parsons' idea to a context in which it will be clear what the "interpretations" are and how they form hierarchies.

Parsons considered the following sentence (p. 227):

(2) The sentence written in the upper right-hand corner of the blackboard in Room 913-D South Laboratory, the Rockefeller University, at 3:15 P.M. on December 16, 1971, does not express a true proposition.

(Note for later use that I shall sometimes abbreviate the sentence-description in (2) as A.)

We are given that sentence (2) was, in fact, written in the upper right-hand corner of the blackboard in the room mentioned on December 16, 1971, and was the only sentence so written. The guiding idea behind Parsons' solution to the Liar is contained on page 230 of his paper, in which he says, in effect (I have slightly simplified the exposition):

(2) says of itself that it does not express a true proposition. Since it does not express *any* proposition, in particular it does not express a true one. Hence it seems to say something true. Must we then say that sentence (2) expresses a true proposition? In either case, we shall be landed in a contradiction. A simple observation that would avoid this is as follows: the quantifiers in one object language could be interpreted as ranging over a certain universe of discourse U. Then a sentence such as

$$(\exists x)(x \text{ is a proposition. } A \text{ expresses } x)$$

is true just in case U contains a proposition expressed by A, that is, by (2). But what reason do we have to conclude from the fact that we have made sense of

(2), and even determined its truth value, that it expresses a proposition that lies *in the universe U*?

It is this rhetorical question that leads Parsons to speak of a hierarchy of interpretations of paradoxical sentences such as (2). To generate a hierarchy of interpretations that can serve as a kind of formal model for what Parsons is suggesting here, I shall begin by using Saul Kripke's idea to generate an initial interpretation.

A Hierarchy Beginning with a Kripkean Interpretation

Saul Kripke is, of course, *the* contemporary logician who put the idea that there is an alternative to Tarski's method, that is, a way to construct a consistently interpreted formalized language so that the truth predicate for the sentences of the language belongs to the language itself, on the map. Since his famous paper,[4] research into that alternative has never ceased. And it should not be surprising that I drag Kripke in here, because sentences such as (II) cannot generate a paradox if we follow Tarski's method; the very fact that they are not in the language that they speak about aborts the paradox. If we wish to formalize the paradoxical reasoning at all, we need a language in which the truth predicate belongs to the language L and not only to a meta-language, and this is the kind of language Kripke showed us how to interpret.

Kripke himself admits that his solution does not wholly avoid hierarchy, for a reason that I shall mention soon, and, of course, the whole moral of Parsons' paper was that, even if the language itself (or natural language itself) is not stratified into object language, meta-language, and so on – that is, even if its *syntax* is not hierarchical – still the best way to think about what is going on with the Liar paradox is to think of a hierarchy: not a hierarchy of formalisms, but a hierarchy of *interpretations* of the syntactically unstratified formalism.

What Kripke achieved was to find a natural way (actually, a whole class of natural ways) to do the following: to assign to the predicate "true" (using a device from recursion theory called "monotone inductive definition") not simply an extension, but a *triple* of sets of sentences, say ⟨Trues, Undecideds, Falses⟩. The first set in the triple – let us call it the Trues – consists of sentences in the object language (henceforth simply L) that are assigned the truth-value "true"; the third set, the Falses, consists of sentences in L that are assigned the truth-value "false"; and the middle set, the Undecideds, contains the remaining sentences, the ones whose truth value is undefined. All this is done in such a way that, of course, the Liar sentence itself turns out to be one of the Undecideds.

I mentioned monotone inductive definition. In fact, the pair ⟨Trues, Falses⟩ is itself the limit of a monotone increasing sequence of (pairs of) sets $\langle \text{Trues}_\alpha, \text{Falses}_\alpha \rangle$, indexed by ordinals; and this whole sequence is a precise mathematical object, as is its limit. Thus, each of the sets Trues, Undecideds, and Falses is itself definable (explicitly and precisely) in a strong enough language, although not in L itself. This is why Kripke speaks of "the ghost of Tarski's hierarchy" as still being present in his construction.

We are going to formally model Parsons' remarks about the Liar paradox in the following way: To simplify matters, we shall not speak of sentences as expressing propositions, as Parsons did in the paragraphs I quoted. Instead, we shall simply think of sentences as true, false, or lacking in truth value. To say of a sentence S that it is not true, that is, to write "$-T(S)$," will be to say something that is (intuitively, as opposed to what happens in Kripke's scheme) *true* if S is either *definitely* false or *definitely* lacking in truth value. In short, "$-T(S)$" says "S is either false or lacking in truth value," or, in Parsons' formulation, "S does not express a true proposition."

We shall assume that, as is usual in formal work, sentences are identified with their Gödel numbers, and that the language L is rich enough to do elementary number theory. Then, by a familar Diagonal lemma, given any open sentence $P(x)$ of the language, we can effectively construct an arithmetical term σ, such that the numerical value of the term σ is the Gödel number of the very sentence $P(\sigma)$, that is, we have a uniform technique for constructing self-referring sentences. Since the language contains the predicate "T" for truth, that is, for truth in an interpretation (although the interpretation will be allowed to vary in the course of our discussion), and hence contains its negation "$-T$," we can effectively find a numerical term τ such that the numerical value of τ is the Gödel number of the very sentence

$$-T(\tau).$$

Speaking loosely, the sentence $-T(\tau)$ says of itself that it is not true or, in Parsons' language, that it does not express a true proposition, and this is precisely the Liar.

We will model Parsons' discussion as follows: we will suppose that when a student – call her Alice – first thinks about the Liar sentence, that is, about the sentence $-T(\tau)$, her first reaction is to say that this is a "meaningless" sentence, that is, not true or false. We shall also follow Parsons by supposing that Alice (implicitly) has an interpretation in mind. Parsons, reasonably enough, supposes that the schemes of interpretation that actual speakers have in mind are only "vaguely defined" (p. 250), but since we are idealizing, we will assume that Alice, implausibly, of course, has in mind precisely one scheme of interpretation, and that it is given by a Kripkean construction.[5] Thus when Alice says of the Liar sentence that it is neither true nor false, she means that it is one of Kripke's Undecideds. But, when we ask her, following Parsons' scenario, whether it follows from the fact that the Liar sentence is neither true nor false that *in particular it is not true*, and we bring her to say "yes," then what has happened (according to Parsons' analysis) – and this seems reasonable – is that she has subtly shifted her understanding (her "interpretation") of the predicate "true" ("T"). The sense in which the Liar, "$-T(\tau)$," is not true is that it does not belong to what we might call the “positive extension” of “T” in Alice’s initial (Kripkean) interpretation, that is, to the set of Trues. She has now shifted to a *bivalent* interpretation of “T” under which “$T(\sigma)$” is true (where σ is any numerical term of language L) just in case the statement that the numerical value of σ lies in the set of Trues is a true sentence of meta-L, the language in which the Kripkean interpretation of L is explicitly defined. Parsons’ own discussion ends at this point; he is content to point out that Alice need not be contradicting herself when she says that the Liar sentence is not true, because the interpretation presupposed by this second remark is not the interpretation of L presupposed by her initial statement that the Liar sentence is neither true nor false. But what interests me is something else – something pointed out some years ago by Professor Ulrich Blau of Munich University, who has not yet published the long work on the paradoxes on which he has been working for many years. What interests me is that the situation is now *unstable*.

Note first that the second interpretation – let us refer to the initial interpretation as Interpretation$_0$ and the second as Interpretation$_1$, henceforth – has a paradoxical feature. For, on the second interpretation, $T(\tau)$ is true just in case the numerical value of the term τ lies in the set of Trues (generated by Interpretation$_0$), and it does not. Hence, $T(\tau)$ is not true, and hence $-T(\tau)$ is true (since Interpretation$_1$ is *bivalent*). But Convention T requires that, if the numerical value of any term σ is (the Gödel number of) a sentence S, then

$$\mathrm{T}(\sigma) \Leftrightarrow \mathrm{S}$$

is true, and τ *is* (the Gödel number of) the sentence $-T(\tau)$. Hence

$$T(\tau) \Leftrightarrow -T(\tau)$$

should be true! Of course, this failure of Convention T is not surprising since, under Interpretation$_1$ ‘T’ does not refer to truth *under Interpretation$_1$ itself* (so Convention T does not really apply) but to truth under the initial interpretation.

The instability, of course, arises because reflection on this new interpretation will generate still another interpretation, and, by iteration, an infinite series of interpretations. To spell this out: under the next interpretation, Interpretation$_{n+1}$, $T(\tau)$ is true just in case the numerical value of the term τ lies in the set of sentences (identified, as we stipulated, with their Gödel numbers) that are true under Interpretation$_n$. Since Interpretation$_1$ is simply the bivalent interpretation of L generated by letting “T” stand for the set of Truths of Interpretation$_0$, and that set is definable in Meta-L, the set of sentences that are true under Interpretation$_1$ is itself definable in Meta-Meta-L (or Meta^2-L). As we have just seen, under Interpretation$_1$, the Liar sentence is true; hence $T(\tau)$ is true under Interpretation$_2$, and hence the Liar sentence is *false* under Interpretation$_2$. In short, the truth value of the Liar sentence flips when we go from Interpretation$_n$ to Interpretation$_{n+1}$, $n > 0$. (Interpretation$_n$ is, of course, definable in Meta^{n+1}-L.)

The series of interpretations can be extended into the transfinite. We shall define a sentence S to be true at a limit ordinal λ if it has *become stably true* at some ordinal $<\lambda$, that is, if there is an ordinal $\kappa < \lambda$, such S is true under Interpretation$_\gamma$ for every γ such that $\kappa < \gamma < \lambda$. Sentences that have become stably true at a stage before λ are true at λ. Similarly, sentences that have *become stably false* at some stage less than λ are false at λ, and sentences that have not become stably true or stably false (e.g., the Liar sentence) are undecided at λ. (Limit interpretations are not bivalent.)

What I want to come to now is the point hinted at in the closing sentences (before the Postscript) of Parsons' paper (p. 251):

> In a simple case, such as that of the word 'I', we can describe a function that gives it a reference, depending on some feature of the context of utterance (the speaker). We could treat the "scheme of interpretation" in this way as argument to a function, but that, of course, is to treat it as an object, for example a set. But a discourse quantifying over *all* schemes of interpretation, if not interpreted so that it did not really capture *all*, like talk of sets interpreted over a set, would have to have its quantifiers taken more absolutely, in which case it would not be covered by any scheme of interpretation in the sense in question. We could produce a "superliar" paradox: a sentence that says of itself that it is not true under any scheme of interpretation. We would either have to prohibit semantic reflection on this discourse or extend the notion of a scheme of interpretation to cover it. The most that can be claimed for the self-applicability of our discussion is that if it is given a precise sense by one scheme of interpretation, then there is *another* scheme of interpretation of our discourse which applies the discourse to itself under the *first* interpretation. But of course this remark applies to the concept "scheme of interpretation" itself. Of it one must say what Herzberger says about truth: in it "there is something schematic . . . which requires filling in."

The sequence of schemes of interpretation of the semantical paradoxes that I just described is a well-defined set-theoretic construction. So far, we have simply associated a scheme of interpretation with each ordinal. (Of course, if we continues it through all the ordinals, then, by cardinality considerations, at some point we will only get interpretations that are extensionally identical to ones already constructed.) But – this is the point that Parsons, citing Herzberger,[6] hints at in the paragraph I just quoted, and the point that Ulrich Blau emphasizes[7] – there is still another source of paradox here. To see this source of paradox, we need to imagine a different scenario than the one Parsons imagined earlier in his paper (our scenario with Alice). There (p. 227 *passim*) Parsons imagines someone who looks at the Liar sentence, decides that it is not true or false (that it is meaningless or, in Parsons' terminology, that it does not express a proposition), and then concludes from that very fact that it is *true* that it doesn't express a true proposition; and he is concerned to argue that that judgment may be totally in order provided we recognize that the scheme of interpretation has changed in the course of the reflection itself. But it seems to me unlikely that this could be the terminus of Alice's reflections. If she is sophisticated, Alice naturally will be led to investigate just the hierarchy of interpretations we constructed, the hierarchy that would result if her act of reflection were iterated through the transfinite.

At this point, a new temptation may arise for Alice, the temptation to land herself in what Parsons refers to as the "superliar paradox." This need not be a temptation to suppose that one can *stand outside* the hierarchy (although one can do that, since the whole inductive definition is carried out, so far at least, within set theory). It is the temptation to suppose that, even *standing within the hierarchy* (and "gazing up," as it were), one can define an *ultimate* sense of "stably true," namely, stably true with respect to the whole hierarchy, and see now that in an *ultimate* sense the Liar sentence is not true (does not express a true proposition), namely that it does not *ever* become stably true. But this, of course, will simply generate a new hierarchy.

Can we go still further? It seems to me that we can. To do so in an interesting way (there are some obvious but uninteresting ways of going further), we will need to use a phrase such as "all the hierarchies one might ever arrive at by continuing reflection," and that means we shall no longer be dealing with precisely defined set-theoretic constructions. This is important, because it may indicate what the answer to my question as to how systematic ambiguity is supposed to differ from just another language might be. When we imagine continuing reflection without limit, creating new hierarchies, and then summing them up – going to the "ultimate interpretation" with respect to a hierarchy, and then taking that ultimate interpretation as the zero stage of a new hierarchy, and so on – we are no longer in the realm of the mathematically well-defined, and hence we cannot assume bivalence (or classical logic). Nevertheless, it does seem that there are things that can be seen to be true in the sense of *provable from the very description of the procedure*. (Compare, in the Tarskian hierarchy, which also can be imagined as extended without limit in similar fashion, the way in which we can see the truth of "For every language L, there is a meta-language *ML* that contains a truth predicate for L.") For example, we see from the very description of the procedure by which any hierarchy is constructed from a given initial interpretation that the Liar sentence never becomes stably true. We cannot imagine an Archimedean point here. We cannot regard the vague "hierarchy of all hierarchies" as something that we can describe *from outside*, as it were. But *we can see from below* how things must go. That is, we can see that *no matter what we "get to"* in the way of reflection on the Liar, no matter what scheme of interpretation we arrive at, we can always use that scheme as the beginning of a new hierarchy, and we can see that, vague as the notion of a hierarchy is, at least *this* much is true of it: the Liar sentence will never become stably true. In fact, using Parsons' device of systematic ambiguity, I can say things like "If the Liar sentence has no truth value at a stage, it gets one to the next stage, and if it has a truth value at a stage, that truth value flips at the next stage." But now it seems to me that Alice may well become the victim of a Super-superliar paradox. The temptation now will be to think something like this:

> When we talk of all the hierarchies of interpretation we could produce, I know that we are not talking about something precise and well-defined, but nevertheless, as you have just shown, there is a sense – a LAST SENSE – in which the Liar sentence is not true: namely, it does not become stably true in *any* hierarchy, not even in, so to speak, the hierarchy of all hierarchies. But surely being eventually stably true in the hierarchy of all hierarchies is the last sense of being stably true, and so there is an absolute sense, namely the LAST SENSE, in which the Liar sentence is not true.

At this point, of course, she will have generated yet another interpretation – an ill-defined one, of course, but nonetheless an interpretation that can also be used as the basis of a hierarchy (even if we have to use intuitionist logic rather than classical logic to talk about it, in view of the fact that the only notion of "truth" that we appear to have in connection with it is some species of provability).

In short, the final temptation is the temptation to suppose that the notion of a LAST scheme of interpretation makes sense. What Parsons says, using a term of Herzberger's, in a sentence I quoted earlier, seems to be exactly right, namely, that when we talk of hierarchies in general, rather than of a specific hierarchy constructed in a specific set-theoretic way, we are necessarily talking schematically; and the schematic character of such talk is, it seems to me, just the difference between talking with systematic ambiguity and merely using Red Ink Language.

There is a further point that I want to make, one emphasized by Ulrich Blau. What is wrong with the temptation to which I said Alice might succumb is not that it is *impossible* to think of an interpretation of the language *L* under which one says of sentences such that we can prove that they will be unstable with respect to every hierarchy that they are "undecided" and of sentences such that we can prove that they will eventually become true in any hierarchy that goes far enough that they are "true" and of sentences such that we can prove that they will eventually become false in any hierarchy that goes far enough that they are "false" – although (because provability is not the same as classical truth) there will be sentences such that we cannot say that they are true, false, *or* undecided if we proceed in this way. As I already mentioned, there are logics that do not assume bivalence (e.g., intuitionist logic), which one might employ in this connection, but I shall not attempt a formal treatment. But if it is all right then, or possibly all right, to treat "the hierarchy of all hierarchies" as something we can reason about, at least in an intuitionist setting, then the mistake that Alice would be making if she gave in to the temptation that leads to what I called the Super-superliar paradox would not be that what I imagined her calling the LAST INTERPRETATION does not *exist*. The mistake is more subtle, it seems to me. The mistake, rather, lies in thinking of it as the "LAST." The phrase "LAST INTERPRETATION" assumes that limits have some kind of finality. But if we allow talk of the LAST INTERPRETATION, we must also allow that there is a *successor* to the LAST INTERPRETATION. That is, it is quite true that the Liar is undecided in the so-called LAST INTERPRETATION, but it is equally true that it becomes true again just AFTER the LAST interpretation. In short, the phrase "LAST INTERPRETATION" is a misnomer. The illusion is that, by this very act of looking up from below at what happens in our hierarchies, we can somehow generate an absolute sense of a "LAST INTERPRETATION," and the paradox itself shows this to be an illusion. Our desire to have a *final thing we can say about the Liar*, or an absolutely best thing to say about the Liar, is what always causes the Liar to spring back to life from the ashes of our previous reflections.

I am led back, in a way, to my own rejected solution in the unpublished James Thompson Memorial Lecture. If you want to say something about the Liar, in the sense of being able to finally answer the question "Is it meaningful or not? And if it is meaningful, is it true or false? Does it express a proposition or not? Does it have a truth value or not? And which one?" then you will always fail. And the paradox itself shows why this desire to be able to say one of these things must always fail.

In closing, let me say that even if Tarski was wrong (as I believe he was) in supposing that ordinary language is a theory, and hence can be described as "consistent" or "inconsistent," and even if Kripke and others have shown that it is possible to construct languages that contain their own truth predicates, the fact remains that the totality of our desires with respect to how a truth predicate *should* behave in a semantically closed language, in particular our desire to be able to say, without paradox, of an arbitrary sentence in such a language that it is true, or that it is false, or that it is neither true nor false, *cannot* be adequately satisfied. The very act of interpreting a language that contains a Liar sentence creates a hierarchy of interpretations, and the reflection that *that* generates does not terminate in an answer to the questions "Is the Liar meaningful or meaningless, and if it is meaningful, is it true or false?" On the other hand, Tarski's own suggestion of giving up on unrestricted truth predicates, and contenting ourselves with hierarchies of stronger and stronger languages, each with its own truth predicate, leaves us in much the same situation as does Parsons' hierarchy of interpretations of a single language. In the end, we are led to see that the things we say about formal languages must be (to use Herzberger's term) "schematic."

NOTES

The present essay and the one that follows it originated as my Alfred Tarski Lectures delivered at the University of California, Berkeley, April 24 and 25, 1995. I thank

Charles Chihara, Doug Edwards, Solomon Feferman, Charles Parsons, and especially Jamie Tappenden for helpful suggestions and criticisms.

1. Charles Parsons, "The Liar Paradox" in his *Mathematics in Philosophy* (Ithaca, NY: Cornell University Press, 1983).
2. "Realism with a Human Face," in H. Putnam, *Realism with a Human Face* (Cambridge, MA: Harvard University Press, 1990).
3. Haim Gaifman, in *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, M. Vardi (ed.), Los Altos, CA: Morgan Kaufman Publishers, 1988, 43–59. "Pointers to Truth," *Journal of Philosophy*, 85 (1992): 223–61.
4. Saul Kripke, "Outline of a Theory of Truth," *Journal of Philosophy*, 72 (1975): 690–716.
5. This may not be as implausible as it sounds. A speaker may well intend that all "paradoxical" sentences be left un-truth-valued, and the preceding might be regarded as one way of rendering that inexact intention precise.
6. Hans Herzberger, "Paradoxes of Grounding in Semantics," *Journal of Philosophy*, 67 (1970): 145–67.
7. In unpublished writing.