KEITH SIMMONS

THE DIAGONAL ARGUMENT AND THE LIAR

I. INTRODUCTION

There are arguments found in various areas of mathematical logic that are taken to form a family: the family of *diagonal arguments*. Much of recursion theory may be described as a theory of diagonalization; diagonal arguments establish basic results of set theory; and they play a central role in the proofs of the limitative theorems of Gödel and Tarski. Diagonal arguments also give rise to set-theoretical and semantical paradoxes. What do these arguments have in common — what makes an argument a *diagonal* argument? And why do some diagonal arguments lead to theorems, while others lead to paradox?

In this paper, I attempt to answer these questions. Cantor's first uses of the diagonal argument are presented in Section II. In Section III, I answer the first question by providing a general analysis of the diagonal argument. This analysis is then brought to bear on the second question. In Section IV, I give an account of the difference between *good* diagonal arguments (those leading to theorems) and *bad* diagonal arguments (those leading to paradox).

The main philosophical interest of the diagonal argument, I believe, lies in its relation to the Liar paradox. The familiar Liar is generated by our ordinary semantical concepts of truth and falsity. Its proper setting is natural language, in which our ordinary semantic terms appear. As Tarski has made clear, this means that the Liar is inextricably linked with another vexed semantical problem, that of universality. Perhaps the central question here is this: Are natural languages universal? Roughly speaking, a language is universal in Tarski's sense if it can say everything there is to be said. If natural languages are universal in this sense, then they can say everying there is to be said about their own semantics. But then it would seem that natural languages fall foul of the Liar.

Journal of Philosophical Logic **19**: 277–303, 1990.

In my view, diagonal arguments are at the heart of the issues raised by the Liar and the problem of universality. In section V, the analysis of good and bad diagonal arguments is applied to a variety of leading solutions to the Liar. I argue that *good* diagonal arguments show the inadequacy of several current proposals. These theories, though quite different in nature, are shown to fail for the same reason: they fail to capture our ordinary semantical concepts. I go on to argue that one version of the claim that natural languages are *not* universal, but expressively incomplete, gives rise to a *bad* diagonal argument, and so leads us back to the Liar. The discussion of Section V provides criteria of adequacy for any solution to the Liar.

II. CANTOR'S USE OF THE DIAGONAL ARGUMENT

In 1891, Cantor presented a striking argument which has come to be known as Cantor's diagonal argument.[1] One of Cantor's purposes was to replace his earlier, controversial proof that the reals are non-denumerable. But there was also another purpose: to extend this result to a general theorem, that any set can be replaced by another of greater power. To these ends, Cantor gave two proofs. The first established the existence of a nondenumerable set which may be associated with the set of reals; the second provided an example of the replacement of a set by one of greater power. Each proof used the method of diagonalization.

The first proof runs as follows. Consider the two elements m and w. Let M be the set whose elements E are sequences $\langle x_1, x_2, \ldots, x_\nu, \ldots \rangle$, where each of $x_1, x_2, \ldots, x_\nu, \ldots$ is either m or w. Cantor asserted that M is nondenumerable, and proceeded to establish this by a proof of the following theorem:

> If $E_1, E_2, \ldots, E_\nu, \ldots$ is any simply infinite[2] sequence of elements of the set M, then there is always an element E_0 of M which corresponds to no E_ν.[3]

Cantor arranged a denumerable list of elements of M in an array:

$$E_1 = \langle a_{1,1}, a_{1,2}, \ldots, a_{1,\nu}, \ldots \rangle,$$

$$E_2 = \langle a_{2,1}, a_{2,2}, \ldots, a_{2,\nu}, \ldots \rangle,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$E_\mu = \langle a_{\mu,1}, a_{\mu,2}, \ldots, a_{\mu,\nu}, \ldots \rangle$$

. . . . . .

Each $a_{\mu,\nu}$ is either m or w. Cantor now defined a sequence $b_1, b_2, b_3, \ldots$, where each of $b_1, b_2, b_3, \ldots$ is either m or w, and, further, if $a_{\nu,\nu} = m$ then $b_\nu = w$, and if $a_{\nu,\nu} = w$ then $b_\nu = m$. Let $E_0 = \langle b_1, b_2, b_3, \ldots \rangle$. Then no E_ν corresponds to E_0. For suppose that $E_0 = E_\nu$, for some ν; then the νth coordinate of E_0 is identical with the νth coordinate of E_ν, which contradicts the definition of the sequence $b_1, b_2, b_3, \ldots$. Notice that this proof may be easily converted into a direct proof of the nondenumerability of the real numbers. If we let $m = 0$ and $w = 1$, then each E_μ is the binary expansion of a real number.

Of his first proof, Cantor wrote

> This proof seems remarkable not only because of its great simplicity, but also because the principle which it follows can be extended directly to the general theorem, that the powers of well-defined sets have no maximum, or, what is the same, that in place of any given set L another set M can be placed which is of greater power than L.[4]

However, Cantor went on to prove not the general theorem, but an instance of it. Cantor took L to be the linear continuum, M to be the set of single-valued functions $f(x)$ which yield only the values 0 or 1 for any value of $x \in [0, 1]$, and proved that M is of greater power than L.[5]

This second proof proceeded in two stages. First, Cantor established that M is at least as large as L, by showing that there is a subset of M which can be put into 1-1 correspondence with L. Consider the following subset of M: the set of those functions on [0, 1] which have value 0 except for one argument x_0. There are as many of these functions as there are reals on [0, 1]. Second, Cantor proved that there is no 1-1 correspondence between M and L. Suppose, towards a contradiction, that there is a 1-1 correspondence ϕ between M and L: for each z in L, there is a function $f(x)$ in M such that $\phi(x, z) = f(x)$, and for each $f(x)$ in M, there is exactly one z in L such that $\phi(x, z) = f(x)$. Now Cantor defined the function $g(x)$ in M, where for any x, $g(x)$ is either 0 or 1, and if $\phi(x, x) = 0$ or 1, then $g(x) = 1$ or 0 respectively. Since $g(x)$ is a single-valued function which yields only the values 0 or 1 for any value of x in [0, 1], $g(x)$ is an element of M. Given the 1-1 correspondence ϕ, there is a z_0 in L

such that $\phi(x, z_0) = g(x)$. Putting $x = z_0$, we obtain $\phi(z_0, z_0) = g(z_0)$; but this contradicts the definition of the function $g(x)$. This completes the second proof.[6]

III. GENERAL ANALYSIS OF THE DIAGONAL ARGUMENT

As we have seen, Cantor's first diagonal argument involves an array, which we may illustrate as follows:

	1	2	3	. . .
E_1	m	w	m	. . .
E_2	w	m	m	. . .
E_3	m	w	w	. . .
⋮	⋮	⋮	⋮	⋱

We can think of this array as composed of two collections — the 'side' ($\{E_1, E_2, E_3, \ldots\}$) and the 'top' ($\{1, 2, 3, \ldots\}$) — and the 'values' m and w. There is a unique value for any pair of elements taken from the side and the top. Any diagonal argument involves such an array.

DEFINITION. Let R be a 3-place relation, and D_1 (the side) and D_2 (the top) be sets.[7] Then, R is an *array* on D_1 and $D_2 \leftrightarrow_{df} \forall x \forall y (x \in D_1 \,\&\, y \in D_2 \rightarrow \exists! z Rxyz)$.

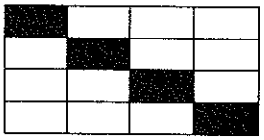
In Cantor's first proof, the array R is given by

$$R(x, y) = \begin{cases} m, \text{ if element } x \text{ has } m \text{ in its } y\text{th place} \\ w, \text{ if element } x \text{ has } w \text{ in its } y\text{th place.} \end{cases}$$

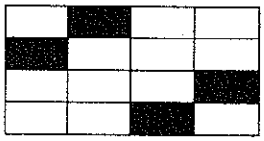
Now consider an array with finite top and side, for example:

	D_2			
D_1	0	1	0	1
	1	1	0	1
	0	0	0	1
	1	1	1	0

Think of the values (here, 0 and 1) as occupying *cells*, given by coordinates $\langle x, y \rangle$, where $x \in D_1$ and $y \in D_2$. We will take diagonals to be composed of cells. The leading diagonal, from top left to bottom right, corresponds to our intuitive notion of a diagonal:



However, other configurations of cells serve just as well in diagonal arguments. What is essential is a 1-1 correlation between D_1 and D_2, and this is equally well supplied by another 'diagonal', say:



In Cantor's presentation of his first proof, for example, the diagonal considered is the leading diagonal, the cells of which are given by $\langle E_1, 1 \rangle$, $\langle E_2, 2 \rangle$, $\langle E_3, 3 \rangle$, ... But the leading diagonal is just one diagonal among many. An alternative is suggested by the coordinates $\langle E_1, 2 \rangle$, $\langle E_2, 1 \rangle$, $\langle E_3, 4 \rangle$, $\langle E_4, 3 \rangle$ . . . . The notion of the leading diagonal does not apply unless top and side are ordered, and there is a correlation corresponding to this ordering. But in diagonal arguments, ordering plays no essential role: it is the correlation between elements of the top and side that is crucial.[8]

So we are naturally led to the following definition of a diagonal.

DEFINITION. F is *a diagonal on* D_1 *and* $D_2 \leftrightarrow_{df} F$ is a 1-1 function from D_1 into D_2.[9]

The notion of a diagonal has to do only with *position*. As yet there is no link between the cells which constitute the diagonal, and the value associated with each cell.

DEFINITION. Let R be an array on D_1 and D_2, and let F be a diagonal on D_1 and D_2. Then G is *the value of the diagonal* F *in* $R \leftrightarrow_{df} \forall x \forall y \forall z (Gxyz \leftrightarrow Fxy \ \& \ Rxyz)$.

In Cantor's first proof, the value in each cell of the leading diagonal is changed. The procedure is illustrated by the replacement of

	1	2	3	...
E_1	m			
E_2		m		
E_3			w	
⋮				⋱

by

	1	2	3	...
E_1	w			
E_2		w		
E_3			m	
⋮				⋱

We introduce the notion of a countervalue in order to generalize this procedure.

DEFINITION. Let R be an array, and F a diagonal, on D_1 and D_2. H is *a countervalue of F in R* $\leftrightarrow_{df}$

(i) $\forall x \forall y(\exists z Hxyz \leftrightarrow Fxy)$

(ii) $\forall x \forall y \forall z \forall z' \, (Hxyz \,\&\, Hxyz' \rightarrow z = z')$

(iii) $\forall x \forall y \forall z \, (Hxyz \rightarrow z \in \text{Range } R)$

(iv) $\forall x \forall y \forall z \, (Hxyz \rightarrow \neg Rxyz)$.

The countervalue corresponding to our illustration of Cantor's first proof may be given as this set of ordered triples:

$$\{\langle E_1, 1, w\rangle, \langle E_2, 2, w\rangle, \langle E_3, 3, m\rangle \ldots\}.$$

Note that if R yields $n + 1$ values ($n \geqslant 1$), there are $n^{\text{card}(\text{dom}F)}$ countervalues. Below, we will make use of this feature of our analysis, that, if R yields more than 2 values, there is more than one countervalue. Previous attempts in the literature to provide a general characterization of the diagonal argument have pointed to this theorem of quantificational logic:

(Ru) $\neg \exists x \forall y(J(x, y) \leftrightarrow \neg J(y, y))$.

The idea is that diagonal theorems are interpretations of (Ru), or some variant of (Ru).[10] Call this the 'Russell analysis'. On the Russell analysis, the analogue of our countervalue is the set of those elements of the domain of discourse which do not bear J to themselves. There

is just one such set. The Russell analysis is captured by a special case of our general analysis, the case where R yields two values, and there is just one countervalue.

In final preparation for the diagonal theorem, we define the notion of a value or a countervalue occurring as a row. A value or countervalue occurs as a row if its associated values form a row of the array.

DEFINITION. Let R be an array on D_1 and D_2, and let K be a value or countervalue of a diagonal F of R. Then, *K occurs as a row of* $R \leftrightarrow_{df} \exists d \in D_1 \forall x \forall y \forall z (Hxyz \rightarrow Rdyz)$.

THE DIAGONAL THEOREM. *Let R be an array on D_1 and D_2 and let F be a diagonal on D_1 and D_2. Let H be a countervalue of F. Then, H does not occur as a row of R.*

Proof.[11]

(1)	Show $\neg \exists w \in D_1 \forall x \forall y \forall z (Hxyz \rightarrow Rwyz)$	
(2)	$\exists w \in D_1 \forall x \forall y \forall z (Hxyz \rightarrow Rwyz)$	Assumption (H occurs as a row)
(3)	$\forall x \forall y \forall z (Hxyz \rightarrow Rdyz)$	2, *EI*
(4)	$\forall x \forall y (\exists z Hxyz \leftrightarrow Fxy)$	Premise (H is a countervalue)
(5)	$\forall y (\exists z Hdyz \leftrightarrow Fdy)$	4, *UI*
(6)	$\forall x \in D_1 \exists y \in D_2 Fxy$	Premise (F is a diagonal)
(7)	$\exists y \in D_2 Fdy$	6, *UI*
(8)	Fde	7, *EI*
(9)	$Fde \rightarrow \exists z Hdez$	5, *QL*, *SL*
(10)	$\exists z Hdez$	8, 9 *SL*
(11)	Hdef	10, *EI*
(12)	Hdef $\rightarrow$ Rdef	3, *UI*
(13)	Rdef	11, 12 *SL*
(14)	$\forall x \forall y \forall z (Hxyz \rightarrow \neg Rxyz)$	Premise (H is a countervalue)
(15)	Hdef $\rightarrow \neg R$def	14, *UI*
(16)	$\neg R$def	11, 15 *SL*

We may distinguish two kinds of diagonal argument: *direct* and *indirect*. In an indirect diagonal argument, the diagonal theorem

is embedded in a proof by *reductio*; in a direct diagonal argument, it is not.

A *direct diagonal argument* specifies in set-theoretical terms a side, a top, an array, and a diagonal, each of which exists. The diagonal result is an interpretation of our diagonal theorem. Cantor's first proof is a direct diagonal argument. The element d appearing in the proof corresponds to Cantor's E_0. By the diagonal theorem, d does not belong to D_1: in Cantor's words, "there is always an element E_0 of M which corresponds to no E_v".

An *indirect diagonal argument* also provides a set-theoretical specification of a side, a top, an array and a diagonal, but assumes the existence of at least one of these *towards a contradiction*. Here, the diagonal argument generates a contradiction, *via* a proof of the diagonal theorem, for the appropriate interpretation. Cantor's second proof is an indirect diagonal argument. Cantor assumed the existence of an array on a side (the set M) and a top (the set L), where the values of the array are 0 and 1. Cantor further assumed, *towards a contradiction*, the existence of a diagonal, the 1-1 correspondence ϕ. Cantor went on to define the function $g(x)$ in terms of ϕ. But here the diagonal theorem tells us that there is no function in M which satisfies the definition of the function $g(x)$, and we have a contradiction.[12]

IV. GOOD AND BAD DIAGONAL ARGUMENTS

Russell remarks that Cantor's diagonal argument

> appears to contain no dubitable assumption. Yet there are certain cases in which the conclusion seems plainly false.[13]

Russell reviews a number of such cases, including his own paradox, and concludes:

> . . . the application of Cantor's argument to the doubtful cases yields contradictions, though I have been unable to find any point in which the argument appears faulty.[14]

As well as those considered by Russell, there are other 'doubtful cases', including Richard's paradox, the heterological paradox, and the cycling and grounding paradoxes in set theory and in semantics. In each of these cases, the diagonal argument leads to a contradiction. And yet in other cases the diagonal argument leads to a theorem.

Why is it that some diagonal arguments are 'good', while others are 'bad'?

Let us begin by considering the bad diagonal argument associated with Richard's presentation of his paradox.[15] Let the side D_1 be the set of real numbers definable by an expression of English, and let the top D_2 be the set of natural numbers. Let $R(x, y) = p$, where p is the digit in the yth decimal place of x. Let F be a 1-1 function from D_1 onto D_2. A countervalue H of F is given by

$$H(x,y) = \begin{cases} p + 1, & \text{if } p \text{ is the digit in the } y\text{th place of the} \\ & \text{decimal expansion of } x, \text{ and } p \neq 8 \text{ or } 9 \\ 1, & \text{if } p \text{ is the digit in the } y\text{th place of the} \\ & \text{decimal expansion of } x, \text{ and } p = 8 \text{ or } 9. \end{cases}$$

The diagonal argument leads to the conclusion that there is no number, definable by an English expression, *which has in its yth decimal place either the number* $p + 1$ *or the number* 1, *according to whether the number correlated with y has in its yth decimal place the number p, where* $p \neq 8$ *or* 9, *or the number* 8 *or* 9. But if we now append the italicized expression in the previous sentence to the expression "The number which has 0 for its integral part and", we obtain an English expression defining a number which, we just concluded, was not definable by an English expression.

Richard offers a solution to his paradox which does not work.[16] But a leading idea of Richard's solution is suggestive. According to Richard, the contradiction is only apparent, because the set that we have labelled D_1 is not 'totally defined'.[17] And it is plausible that the problem lies with D_1. Suppose that we had started out by assuming, towards a contradiction, that there is a set D_1 of reals definable by English expressions. We could then have proceeded with the diagonal argument of the previous paragraph, and obtained a contradiction. Now, instead of a paradox, we have a *reductio* proof that there is no such set D_1. We obtain an indirect diagonal argument. According to this response to Richard's paradox, there is nothing wrong with the *reasoning* of the bad diagonal argument — rather, what is at fault is the assumption that all of the top, side, array and diagonal exist. This bad diagonal argument, unlike direct diagonal arguments, assumes

the existence of a set that does not exist. And unlike indirect diagonal arguments, this assumption is not made towards a contradiction.[18]

There are good direct analogues of this bad diagonal argument. If we let D_1 be any denumerable set of reals, and keep the rest of the interpretation fixed, we obtain a direct diagonal argument which constitutes a proof of the non-denumerability of the reals, quite analogous to Cantor's first proof.

A similar diagnosis can be made for other bad diagonal arguments. Consider Russell's paradox. Suppose we assume that there is a set M of exactly those sets which are not members of themselves. The side and the top are each the proper class of all sets, *where the side and top are taken to include the set M*. The array R is given by

$$R(x, y) = \begin{cases} 1, & \text{if } y \in x \\ 0, & \text{if } y \notin x. \end{cases}$$

The diagonal F is identity. The countervalue H of F is given by

$$H(x, x) = \begin{cases} 1, & \text{if } x \notin x \\ 0, & \text{if } x \in x. \end{cases}$$

By the diagonal theorem, there is no set of exactly those sets which do not belong to themselves, contradicting our assumption. It is the assumption that the set M exists that generates the paradox. If we make this assumption for *reductio*, we obtain an indirect diagonal argument.

Again, there are good analogues of this bad diagonal argument. Consider, for example, the argument from recursion theory which establishes the recursive unsolvability of the halting problem. A basic theorem on recursively enumerable sets states: A is r.e. iff A is the domain of a partial recursive function. Let W_x = domain ϕ_x, where x is a Gödel number for the r.e. set W_x. Let $K = \{x \mid \phi_x(x) \text{ convergent}\} = \{x \mid x \in W_x\}$. So $\bar{K} = \{x \mid x \notin W_x\}$. One way of expressing the recursive unsolvability of the halting problem is to say that K is not recursive. This we can prove by showing that $\bar{K}$ is not r.e., since K is r.e., and, in general, A is recursive iff A and $\bar{A}$ are both r.e. The proof that $\bar{K}$ is not r.e. is a diagonal argument. The side D_1 and the top D_2 are a set

of Gödel numbers of all r.e. sets. The array R is given by

$$R(x, y) = \begin{cases} 1, & \text{if } y \text{ is a member of the r.e. set with Gödel no. } x \\ 0, & \text{if } y \text{ is not a member of the r.e. set with Gödel no. } x. \end{cases}$$

The diagonal F is identity. The countervalue H of F is given by

$$H(x, x) = \begin{cases} 1, & \text{if } x \text{ is not a member of the r.e. set with Gödel no. } x \\ 0, & \text{if } x \text{ is a member of the r.e. set with Gödel no. } x. \end{cases}$$

By the diagonal theorem, H cannot occur as a row. So the set $\bar{K} = \{x \mid x \notin W_x\}$ is not r.e. Clearly, this diagonal argument is analogous to that associated with the Russell paradox. But here the direct diagonal argument is good. Unlike M, the set $\bar{K}$ exists.[19]

A bad diagonal argument may take the form of an *indirect* argument. We can generate Cantor's paradox as follows. Let the top D_2 be the set of all sets, and let the side D_1 be the power set of D_2. The array R is

$$R(x, y) = \begin{cases} 1 & \text{if } y \in x \\ 0 & \text{if } y \notin x. \end{cases}$$

We suppose towards a contradiction that there is a diagonal F on D_1 and D_2. We can now define a certain member of D_1 in terms of F: the set of those elements of D_2 that do not belong to the subset of D_1 with which they are correlated by F. But by the diagonal theorem, there is no such subset in D_1. We have a contradiction, and so there is no diagonal F. And now we have a paradox: D_1 is neither smaller nor equal in size to D_2, yet D_2 is the set of all sets. This argument is bad because it assumes the existence of a universal set. It is readily converted to a proof that there is no universal set. And there is a good analogue of this bad diagonal argument: let the top be any set, and we obtain the indirect diagonal argument that establishes Cantor's power set theorem.

Finally, let us turn to a bad diagonal argument related to the Liar. We introduce into English the 1-place predicate 'heterological', denoting the property of being a predicate that does not have the property it denotes. (So, for example, the predicates 'French', 'monosyllabic', and 'ambiguous' are heterological.) Now let the side and the top be the set of 1-place predicates of English, *where the side and top are taken to include the predicate 'heterological'*. The array R is given by

$$R(x, y) = \begin{cases} 1, & \text{if } y \text{ has the property denoted by } x \\ 0, & \text{if } y \text{ does not have the property denoted by } x. \end{cases}$$

The diagonal F on R is identity. The countervalue H of F is given by

$$H(x, x) = \begin{cases} 1, & \text{if } x \text{ does not have the property denoted by } x \\ 0, & \text{if } x \text{ does have the property denoted by } x. \end{cases}$$

By the diagonal theorem, there is no predicate of English that denotes the property had by exactly those predicates of English that do not have the property they denote; and yet 'heterological' is just such a predicate. We are landed in paradox.

The conversion of this bad diagonal argument to an indirect diagonal argument constitutes an attempt to solve the paradox. We might take the assumption for reductio to be the assumption that there is a 'bivalent' array, that the concept of heterologicality is everywhere applicable — allowing truth value gaps is one approach along these lines. An alternative response is to say that though there is such a concept, there is no predicate of the language that expresses it — we assume towards a contradiction that the top and side include a predicate expressing this concept. Both these responses will be discussed below.

There are good direct analogues of this bad diagonal argument. As Tarski remarks,[20] there is a close analogy between the heterological paradox and Tarski's theorem about truth. We capture Tarski's theorem as follows. Let S be a usual first-order theory with identity which is based on Peano's postulates and is adequate for the proofs of all the basic results of number theory. The side D_1 is the set of

1-place wffs of S, and D_2 is the set of natural numbers. The array R is given by

$$R(x, y) = \begin{cases} 1 & \text{if } x \text{ is true (in the standard model) of } y \\ 0 & \text{if } x \text{ is not true of } y. \end{cases}$$

The diagonal F carries each wff in D_1 to its Gödel number. The countervalue H of F is

$$H(x, y) = \begin{cases} 1 & \text{if } x \text{ is not true of its Gödel number } y \\ 0 & \text{if } x \text{ is true of its Gödel number } y. \end{cases}$$

By the diagonal theorem, no wff of S is true of exactly the Gödel numbers of wffs not true of their own Gödel numbers. But there is such a wff of S on the assumption that the set of Gödel numbers of wffs of S that are true in the standard model is arithmetical. We conclude that this set is not arithmetical. The diagonal argument here closely resembles that of the heterological case. But the set of wffs of S not true of their own Gödel numbers exists; and so too does a wff with this set as its extension, in a suitable metalanguage. There is nothing problematic about the set or the wff; this is in contrast with the predicate 'heterological' and the concept of heterologicality.

To sum up, bad diagonal arguments specify a top, a side, an array, and a diagonal, in set-theoretical terms. As with indirect diagonal arguments, the specification of this set-theoretical apparatus involves somewhere the assumption of a non-existent set. And so, like indirect diagonal arguments, bad diagonal arguments generate a contradiction, via valid reasoning that incorporates the proof of the diagonal theorem. Unlike indirect diagonal arguments, this contradiction is not part of a proof by *reductio*. The conversion of a bad diagonal argument to a good indirect one may be a straightforward matter. But not always; as we will now see, the lesson of the Liar is not obvious.

V. THE DIAGONAL ARGUMENT AND THE LIAR

Tarski writes:

> A characteristic feature of colloquial language (in contrast to various scientific languages) is its universality. It would not be in the spirit of this language if in some other language

a word occurred which could not be translated into it; it could be claimed that 'if we can speak meaningfully about anything at all, we can also speak about it in colloquial language'.[21]

In particular, says Tarski, natural language

is semantically universal in the following sense. Together with the linguistic objects, such as sentences and terms, names of these objects are also included in the language . . . ; in addition, the language contains terms such as "truth", "name", "designation", which directly or indirectly refer to the relationship between linguistic objects and what is expressed by them.

In what follows I shall be concerned with universality and semantic universality in Tarski's sense.[22]

Some truth-value gap theorists are motivated by the intuition that natural languages are universal. According to R. L. Martin, gap solutions of the kind he endorses "retain the intuitive view of language as universal and give up intuitions about what we thought there was to be said".[23] Faced with the heterological paradox, we may be pulled in two different directions. We might be drawn to a gap approach like Martin's. According to such a view, sentences like "'Heterological' is heterological" are without truth value, and the associated paradox-producing reasoning is blocked. On the other hand, we might convert the heterological paradox into a proof that *heterologicality* is an inexpressible concept, since the assumption that there is a term which expresses this concept leads to contradiction. According to this alternative view, natural language is not universal, but expressively incomplete.

But now Martin claims that if we allow truth-value gaps, we may deny that there is any such concept of heterologicality *to be* expressed. The gap at the level of language is matched by a gap at the level of ontology. Martin speaks of another gap, between the situations before and after analysis: before analysis we thought there was more to be said than what analysis reveals there is to be said. There is only a *restricted* concept of heterologicality to be expressed. We might think of it as a Fregean concept which yields neither truth nor falsity as values for certain arguments.

If we accept Martin's argument thus far, hopes of universality are not yet dashed. However, given truth-value gaps, we can legitimately form the concept *is an expression which is false or neither true nor false of itself* (or *superheterologicality*, for short). The assumption that

this concept is expressible leads to a contradiction. For suppose that the term 'superheterological' denotes this concept. Then, whether we assume 'superheterological' is true of itself, false or itself, or neither true nor false of itself, we obtain a contradiction. Adopting gaps and assuming universality leads to contradiction: the gaps allow the construction of concepts which, if assumed to be expressible, generate paradoxes. And these new paradoxes arise out of the appeal to gaps, and must be resolved in some other way. So the point is not just that an appeal to truth-gaps fails to preserve intuitions about universality. The truth-value gap theorist fails also to provide a general, unified account of semantical paradox.

This is a criticism that can be made of Kripke's truth-gap approach to the Liar.[24] By a fixed point construction, Kripke obtains a language $\mathscr{L}_\sigma$ which expresses its own concept of truth; that is, there is a formula of $\mathscr{L}_\sigma$ which is true of exactly the codes of true sentences of $\mathscr{L}_\sigma$. However, it can be shown that the complement of truth is *not* expressible in $\mathscr{L}_\sigma$.[25] The argument that yields this result is a diagonal argument. Let the side D_1 be the set of 1-place wffs of $\mathscr{L}_0$ and let the top D_2 be the set of natural numbers. The array R is given by

$$R(x, y) = \begin{cases} 1, & \text{if } x \text{ is true of } y \\ 0, & \text{if } x \text{ is not true, i.e. is false or undefined, of } y. \end{cases}$$

The diagonal F is a 1-1 function which carries each wff in D_1 to its code. The countervalue H of F is given by

$$H(x, y) = \begin{cases} 1, & \text{if } x \text{ is not true (is false or undefined) of its code } y \\ 0, & \text{if } x \text{ is true of its code } y. \end{cases}$$

The diagonal theorem tells us that no wff of $\mathscr{L}_\sigma$ is true of exactly the codes of those wffs false or undefined of their own codes. This is an exact analogue of Tarski's theorem, with 'not true' now understood as 'false or undefined', since we are admitting truth gaps. Just as Tarski's good diagonal argument is associated with the heterological paradox, so our good diagonal argument is associated with the

superheterological paradox. Both arguments establish that the respective object languages are not semantically universal.

Our theorem indicates that Kripke's truth-value gap approach cannot dispense with a Tarskian hierarchy.[26] The theorem forces the first step up such a hierarchy, to a metalanguage for $\mathscr{L}_\sigma$, in which we can talk about the complement of truth.[27] We are forced up the hierarchy in order to avoid semantical paradoxes with which truth gaps cannot deal. And this shows that Kripke's gap theory does not provide a *general* solution to the Liar: ultimately, it is a Tarskian hierarchy of languages that allows us to escape semantical paradox.[28]

I think Kripke must (and in fact does) accept this objection. And I think Kripke would respond by arguing that the limitations forced on his theory do not diminish its significance. According to Kripke, though the minimal fixed point does not model a universal language, *it is a model of a significant stage of development of natural language*; it is a model of "natural language at a stage before we reflect on the generation process associated with the concept of truth, the stage which continues in the daily life of nonphilosophical speakers".[29] Kripke is claiming that, *for this stage of natural language*, his theory provides an adequate account of truth.[30]

Let us return to the proof of our theorem. Informally, we construct the sentence '"Is not true of itself" is not true of itself', and go on to derive a contradiction; the theorem requires only the notions of truth and negation. The relevant notion of negation here is *exclusion negation*. If we interpret "$\neg$" as exclusion negation, $\neg A$ is true iff A is false or undefined, and $\neg A$ is false iff A is true. Exclusion negation is contrasted with *choice negation*. If "$\neg$" is interpreted as choice negation, $\neg A$ is true iff A is false, $\neg A$ is false iff A is true, and $\neg A$ is undefined iff A is undefined. Now, Kripke's construction of the minimal fixed point uses choice negation: our theorem shows that the construction cannot be carried out if negation is taken to be exclusion negation. So Kripke relegates exclusion negation to a metalanguage.[31] Others too have sought to defend truth-value gap approaches this way. Terence Parsons writes:

> When we "exclude exclusion negation" from our language we are not in fact excluding anything at all. For there is no such thing as exclusion negation in any formal language which accurately reflects our own native speech.[32]

This claim seems to me wrong as a matter of empirical fact.[33] But I think that this focus on negation is misguided anyway. It is not *negation* that is the real source of the gap theorist's troubles. Let us construct the superheterological paradox according to our analysis of the diagonal argument. The side D_1 and the top D_2 are the 1-place predicates of English, the array R is given by

$$R(x, y) = \begin{cases} t, & \text{if } x \text{ is true of } y \\ f, & \text{if } x \text{ is false of } y \\ u, & \text{if } x \text{ is neither true nor false of } y, \end{cases}$$

and the diagonal F is identity. There are 2^{ω} ways of forming a countervalue (recall the remark on p. 282). And just one of these countervalues, call it H_N, is associated with exclusion negation:

$$H_N(x, x) = \begin{cases} f, & \text{if } x \text{ is true of } x \\ t, & \text{if } x \text{ is false of } x \\ t, & \text{if } x \text{ is neither true nor false of } x. \end{cases}$$

This countervalue builds the concept *false of itself or neither true nor false of itself*, which can be alternatively expressed as *not true of itself*, where 'not' is exclusion negation. But exclusion negation is not required to express this concept: we have just constructed it from the notions of *false* and *neither true nor false*. And the concepts associated with the other countervalues are each expressible in terms of the notions of *true*, *false* and *neither true nor false*.[34] It is, in general, a mistake to see the emergence of paradox as having anything essentially to do with exclusion negation: *what is essential, rather, is the construction of a countervalue*.

In the present case, each countervalue is constructed from notions which *are* to be found in 'nonphilosophical' language. In particular, the notion of a sentence being neither true nor false surely is in the repertoire of the ordinary speaker. This notion is composed of the everyday notions of truth and falsity, and the 'neither-nor' construction. And the notion has clear, intuitive application — to meaningless or nonsensical sentences, for example. Indeed, gap theorists themselves motivate truth

gaps by appeal to our semantic intuitions, in ways independent of semantical paradox. For example, Martin appeals to category considerations, while van Fraassen starts out from the Frege–Strawson theory of presupposition. Kripke himself is motivated by Strawson's doctrine. Of course, defending the introduction of truth gaps in these ways will involve *some* semantic reflection on language. But such reflection is couched in quite ordinary language; to use Kripkean terminology, it is expressed in language at a stage *prior to* reflection on the generation process associated with the concept of truth. Such intuitive motivation does not involve philosophical reflection on the Liar.

These are points missed by those who would defend truth-value gap theories by excluding exclusion negation. Parsons argues that though the non-creative definition

$$\psi \ =_{df} \text{ the function which maps } t \text{ to } f \text{ and both } f \text{ and } u \text{ to } t$$

defines a function, the existence of exclusion negation is not thereby guaranteed. For, according to Parsons, it must also be the case that "the truth-function in question can be assigned as the denotation of a unary connective that consistently forms falsehoods from truths and truths from sentences that are either false or neuter".[35] But even if there is no such unary connective in ordinary English, there are other means available *within ordinary English* for the construction of the associated countervalue, and other countervalues too. Each of these countervalues is associated with a version of the Liar which may be expressed in ordinary language, but may not be resolved by an appeal to truth-gaps.[36]

Let me recapitulate my objection to Kripke's theory. There are two stages. First, our diagonal theorem demonstrates the need to ascend to a metalanguage: hence, Kripke's truth gap approach is not a full solution to semantical paradox. Second, Kripke cannot retreat to the claim that his approach provides a solution to the Liar for a certain significant stage of natural language: diagonal arguments show that this stage of natural language has the resources to formulate paradoxes which Kripke's theory cannot resolve.

This two-part objection to Kripke's theory may be generalized. *In its general form, it presents a challenge to any purportedly non-Tarskian*

approach to the Liar, not just those, like Kripke's, which appeal to truth-gaps. We can express the challenge in the form of two questions. First, does the theory give rise to semantical concepts which can be expressed only in a metalanguage, on pain of paradox? If the answer to this question is affirmative, then the scope of the proposed non-Tarskian theory is limited. And an affirmative answer to the first question prompts this second question: Are these semantical concepts available to the ordinary speaker, independently of philosophical reflection on the Liar? If the answer to this second question is also yes, then not only is the scope of the proposed solution put into question, but so is its significance. For, to repeat what was said at the outset, the Liar is a product of our ordinary semantical concepts, expressed by our ordinary semantical terms. A proposed solution that fails to give an account of our semantical concepts fails to come to grips with the Liar.

Gupta and Herzberger offer similar modifications of Kripke's theory which admits the classical valuation scheme: the 'anti-extension' of the truth predicate complements its extension.[37] But now a key notion of their theories, the notion of *stable truth*, is a notion of the metalanguage, as an indirect diagonal argument will establish.[38] So the answer to the first question is affirmative.

Gupta and Herzberger suggest that the stable sentences of the formal object language L capture our intuitive notion of *semantically unproblematic sentences*.[39] But now the paradox of stable truths may be expressed in intuitive terms. We can form the concept associated with the relevant countervalue, the concept *does not yield a semantically unproblematic truth when appended to its own quotation.* Paradox issues in the usual diagonal fashion. This is a version of the Liar expressible in ordinary terms, but beyond the reach of Herzberger and Gupta's theory, as presented. The answer to our second question is also affirmative.

Gupta suggests, without elaboration, that we may add the predicate 'stably true in L' to the language L. Now

> the paradox is present for the concept "stably true in L". But we must ask how is the concept "stably true in L" added to L? It must be added, it would appear, by a rule of revision. But then can we not give an account of the new paradox parallel to that we gave of the old?[40]

Now the following sentence is a sentence of (L):

(S) (S) is not stably true in L.

Since (S) is paradoxical, (S) is not a stable truth of (L). That is, we may assert:

$(S)'$ (S) is not stably true in L.

So, while (S) is paradoxical, $(S)'$ is a true assertion. To account for this, we need to distinguish two stable truth predicates, one internal to L, and one external to L. The internal stable truth predicate is in the object language, and expresses the concept for which the paradox is present. The second is in a metalanguage for L, and (S) is never in its extension. The need for the essentially richer language may be demonstrated by a good diagonal argument.

We might yet follow Gupta's general line by enriching the object language with this external stable truth predicate, expressing a further concept added by a rule of revision. But still this enriched object language will not express *its* external notion of truth. There is no end to this series of increasingly rich object languages. No language in the series will express its external notion of stable truth: a good diagonal argument guarantees this. This series of languages is quite analogous to a Tarskian hierarchy in which each language includes its predecessor as a proper part.[41]

Rescher and Brandom present an altogether different kind of solution to the Liar.[42] According to their view, Liar sentences are both true *and* false. Theirs, then, is an *inconsistency* view, a view which has received increasing support in recent years. However, by their own admission, although their theory can handle the Liar sentence "This sentence is false", it *cannot* handle the Liar sentence "This sentence is not true".[43] According to Rescher and Brandom, we must separate the inconsistent object theory from our consistent discourse about it.[44] But this suggests a way out of the Liar which is ultimately along Tarskian lines.[45] And the inconsistency approach itself leaves untreated perfectly ordinary versions of the Liar. For this inconsistency approach, we have an affirmative answer to both our questions.[46]

The approaches to the Liar that I have considered in this section have failed in the same general way. In each case, a good diagonal

argument demonstrates that the object language is not semantically universal. Perhaps, then, *this is just the lesson that the Liar teaches*: there are (semantical) concepts which natural language cannot express.[47] For the assumption that a certain semantical concept can be expressed by a term of (say) English leads to a contradiction, associated with some version of the Liar.

Such a line is taken by Herzberger.[48] Herzberger's inexpressibility claims take the following form: a concept is inexpressible in some conceptual system if the semantic rules assign its extension to no term of that system.[49] For example, Herzberger argues that the set of heterological terms of English is the extension of no term of English;[50] and that no term of English has as its extension the set of grounded terms of English.[52] In each case, Herzberger assumes for *reductio* that there is a term of English with the given set as extension, and obtains a contradiction associated with semantical paradox.

I shall now argue that we should reject these inexpressibility claims.[52] Consider an extensional version of the heterological array on p. 288. If we follow Herzberger's line, we have no Liar-related reason to suppose that we are unable to fill in all the values of the array. Notice that though the English expression 'heterological' (or, 'is not true of itself') is in the side and top, there is no particular problem about its extension, since, according to Herzberger's claim, whatever its extension is, it is not such as to produce semantic paradox. So it is a consequence of Herzberger's claim that the array can be completed. Now we can produce a countervalue in the usual way. And associated with this countervalue is a certain set (*the set of heterological terms of English*, we might be tempted to say, but of course these italicized words won't do the job). Our analysis provides a precise way of specifying the countervalue, given the fully determinate array, and so a precise way of specifying the associated set of English predicates.

But since our analysis may be expressed in English, the set associated with the countervalue may be specified in English. To deny this is to deny that we can talk about what we clearly *can* talk about: we surely can talk about the array, and the various functions and sets associated with it. The countervalue is a determinate set of ordered triples, and the associated set of predicates of English is a determinate

set of English predicates definable in terms of the countervalue. And we can talk about all this in English — indeed, that is just what we are doing. Herzberger's inexpressibility claim has itself provided conditions that allow the specification of the countervalue and the associated set, *in English*. But if can specify this set in English, then the set is the extension of a predicate of English.

Let '*Het*' stand for this predicate. Is *Het* a member of the top and side of our array? If it isn't, then neither the side nor the top is the set of *all* 1-place English predicates, *contra* Herzberger's assumption. Pursuing this line, one might conclude that a natural language like English is indefinitely extendible, continually expanding to cover more concepts. But this is not the thesis that Herzberger is proposing.

Suppose, then, that *Het is* a member of the top and side. Now, of course, we are landed in contradiction. In assuming that *Het* is a member of the top and side, we assume that we can fill in the values of its row and column; that is, we assume it has a definite extension. But supposing that we can fill in the value for the cell $\langle Het, Het \rangle$ leads to a contradiction. Herzberger does not prevent the formation of an array associated with a bad diagonal argument. The thesis of expressive incompleteness does not escape paradox.[53] It will do no good to conclude that the set associated with the countervalue is the extension of no predicate of English, not even *Het*, for the argument is re-iterable. And if we say there is no such set, then we have given up Herzberger's thesis of expressive incompleteness.

Good diagonal arguments have shown that certain purportedly non-Tarskian theories of truth ultimately cannot dispense with a Tarskian hierarchy. This limits the scope of the theories. Further, since these diagonal arguments utilize ordinary semantical concepts, the significance of the theories, *qua* solutions to the Liar, is questionable. We can extract here a simple criterion of adequacy for a solution to the Liar: the formal theory must represent our semantical terms and concepts. I have argued that the theories discussed in the course of this paper fail to meet this criterion.[54]

We may identify another related criterion. A solution to the Liar must do justice to the expressive capacity of natural language. Tarski's intuition that natural languages are universal is not easily

dismissed. We should beware of claims that natural language is not semantically universal, that certain semantical concepts are inexpressible. As we have seen, at least one claim of this sort gives rise only to bad diagonal arguments and paradox.[55,56]

NOTES

[1] Georg Cantor, "Über eine elementare Frage der Mannigfaltigkeitslehre", first published in *Jahresbericht der Deutschen Mathematiker-Vereinigung* I (1890–91), pp. 75–78, and reprinted in *Gesammelte Abhandlungen mathematischen und philosophischen Inhalts*, ed. E. Zermelo, Berlin: J. Springer 1932, pp. 278–281. All page references are to the latter collection.

[2] A 'simply infinite' sequence is a denumerable sequence, one that can be put into 1-1 correspondence with the natural numbers.

[3] Cantor 1891, p. 278. I follow the translation in J. W. Dauben, *Georg Cantor*, Harvard University Press 1979, p. 165.

[4] Cantor, *op. cit.*, p. 279.

[5] Strictly, this is to claim too much. To show that M is of greater power than L, it must be shown that there is no 1-1 function from L into M; and for this the Shröder-Bernstein theorem is needed. No proof of this existed in 1891 when Cantor wrote his paper.

[6] Cantor's second proof is easily converted into a proof that the linear continuum is less in power than the set of all its subsets, and a straightforward extension of this latter proof establishes Cantor's general theorem.

[7] This definition is easily extended to proper classes.

[8] There is another way in which we generalize the notion of a diagonal: we extend it to nondenumerable arrays. For an array in which one or both of the side and top are finite, there is a clear sense to the intuitive notion of the diagonal. For an array in which the side and top are either finite or denumerably infinite, sense still attaches to the notion of the diagonal, though it is now understood as infinitely extendible. But where either side or top is nondenumerable, as in Cantor's second proof, the intuitive notion of the diagonal breaks down.

[9] According to this definition, a diagonal passes through every row, but not necessarily every column. One kind of diagonal passes through every row and every column: in his proofs, Cantor uses this kind of diagonal.

[10] It is usually J. F. Thomson who gets the credit for pointing out the connection between this theorem, various paradoxes and the diagonal argument, in his paper, "On Some Paradoxes", in R. J. Butler (ed.), *Analytical Philosophy*, Oxford 1962. But Russell had, in effect, made the connection; see *The Principles of Mathematics*, pp. 366–368, esp. Section 347. Herzberger uses (Ru) and variants to analyze a family of diagonal arguments, in "Paradoxes of Grounding in Semantics", *Journal of Philosophy* 1970. This theorem is also discussed in R. L. Martin, "On a Puzzling Classical Validity", *Philosophical Review* 1977, and in Leonard Goddard and Mark Johnston, "The Nature of Reflexive Paradoxes: Part I", *Notre Dame Journal of Formal Logic*, 1983.

[11] The proof is largely in the style of Kalish, Montague and Mar, *Techniques of Formal Reasoning*, Second Edition, Harcourt Brace Jovanovich, 1980.

[12] Direct and indirect diagonal arguments are the two kinds of diagonal arguments that establish theorems (these are the 'good' diagonal arguments). In the next section we shall also characterize 'bad' diagonal arguments, which are neither direct nor indirect diagonal arguments.

[13] Russell, *op. cit.*, p. 366.

[14] *op. cit.*, p. 368.

[15] See Jules Richard, "Les principes des mathématiques et le problème des ensembles", in *Revue générale des sciences pures et appliquées*, 1905. Richard's paper is translated in Jean van Heijenoort (ed.), *From Frege to Gödel*, Harvard University Press 1967, pp. 143–144.

[16] Richard (*op. cit.*) not only presents his paradox, but goes on to offer a solution. In a paper in preparation, I argue that the most plausible reading of Richard's brief comments do not block the re-emergence of the paradox.

[17] Van Heijenoort, *op. cit.*, p. 143.

[18] Of course, an adequate solution to Richard's paradox would provide an account of the notion of *definability* that would explain *why* there is no such set.

[19] By the definition of K, we have that $W_x C \bar{K} \rightarrow x \in \bar{K} - W_x$. This property of $\bar{K}$ is given a recursively invariant formulation in the definition of a *productive* set. Productiveness is closely linked to diagonalization: theorems concerning productive sets provide further examples of diagonal arguments.

Other basic theorems of recursion theory would have served our illustrative purposes just as well; for example, the proof that the class of primitive recursive functions does not include all algorithmic functions, or Kleene's result that there is no algorithm which yields just the total functions.

[20] See Tarski's "The Concept of Truth in Formalized Languages", in *Logic, Semantics, Metamathematics*, Oxford 1956, p. 248, fn. 2.

[21] Tarski, *op. cit.*, p. 164.

[22] It is worth pointing out that Tarski is not claiming that all concepts are expressible in, say, English. Such a claim would be trivially false. Rather, Tarski is saying that if a concept is expressible in some language, then it is expressible in English. And the narrower claim that English is semantically universal is the claim that English has the means to express its semantic concepts, such as the concepts of truth, falsity, and reference. Below, I shall add to this list the semantic concepts *neither true nor false* and *semantically unproblematic sentence*.

[23] R. L. Martin, "Are Natural Languages Universal?", *Synthese* 1976, p. 288.

[24] Saul Kripke, "Outline of a Theory of Truth", *Journal of Philosophy* 1975, reprinted in R. L. Martin (ed.), *Truth and the Liar Paradox*, Oxford 1984. All page references are to the Martin anthology.

[25] Elsewhere, I have presented Kripke's theory, and the result just mentioned, in a fully rigorous way, using Moschovakis's notion of an acceptable structure.

[26] Kripke is aware of the necessity to ascend to a hierarchy, but perhaps understates the problem when he writes: "The ghost of the Tarski hierarchy is still with us" (*op. cit.*, p. 80).

[27] One such metalanguage (call it **M**) is the language of Kripke's paper. According to Kripke, **M** can be regarded as containing no truth gaps, since a sentence either does or does not have a truth value in a given fixed point. But now Tarski's theorem applies to **M**: "true-in-**M**" is not contained in **M**, but in some further metalanguage. And so a Tarskian hierarchy is generated. Alternatively, one could add to **M** a predicate T', and

by Kripke's construction obtain a fixed point interpretation of $\mathbf{M} + T'$, so that T' is the truth predicate of $\mathbf{M} + T'$. Then we will need a further metalanguage to express the complement of T' — and again, a Tarskian hierarchy is generated.

[28] Tyler Burge, in "Semantical Paradox", *Journal of Philosophy* 1979, also objects to Kripke's theory along these lines.

[29] *Op. cit.*, fn. 34, p. 80.

[30] Perhaps we should be suspicious of this talk of a certain stage of natural language; see Tyler Burge, 'Semantical Paradox', *Journal of Philosophy*, 1979, p. 88, fn. 9. But I shall not pursue such worries in this paper.

[31] *Op. cit.*, p. 80, and fn. 35.

[32] Terence Parsons, "Assertion, Denial, and the Liar Paradox", *Journal of Philosphical Logic* 1984, p. 149.

[33] The notion of a meaningless sentence (in its ordinary, non-technical sense) surely is in the repertoire of nonphilosophical speakers, and it is natural enough to infer 'A is not true' from '*A* is meaningless'. And here is a use of 'not' which is most plausibly analyzed as exclusion negation. Further, if we accept that ordinary speakers have available to them the notion of a sentence being neither true nor false, then the inference from '*A* is neither true nor false' to '*A* is not true' is an intuitive one, and a use of exclusion negation appears in the conclusion.

[34] For example, the concept corresponding to the countervalue

$$H(x, x) = \begin{cases} u & \text{if } x \text{ is true of } x \\ t & \text{if } x \text{ is false of } x \\ f & \text{if } x \text{ is undefined of } x \end{cases}$$

is expressed by *undefined of those expressions true of themselves, true of those expressions false of themselves, and false of those expressions undefined of themselves.*

[35] Parsons, *op. cit.*, p. 150.

[36] The limitations of truth-value gap approaches to the Liar might suggest that we haven't gone far enough, that we should make a more radical break with classical logic and semantics. Here, fuzzy logic might seem a natural candidate. The term 'fuzzy logic' is sometimes applied to systems which admit *non-denumerably many* truth-values. By a generalization of Lukasiewicz's 3-valued logic, these truth values may be identified with the real numbers in the interval [0, 1]. And other approaches are possible. According to Zadeh, the extended Lukasiewicz system is a *non*fuzzy multi-valued system, and should be regarded only as a basis for a fuzzy logic. Zadeh starts with the non-denumerable set of truth values and from it constructs truth values that are themselves fuzzy (see Zadeh's "Fuzzy Logic and Approximate Reasoning", *Synthese*, 1975).

But on any 'fuzzy' approach, we obtain a 'fuzzy' analogue of the array associated with the superheterological paradox. Associated with this array are infinitely many countervalues. These generate infinitely many inexpressible concepts, and infinitely many versions of the Liar that fuzzy logic cannot accommodate. Clearly, we are no better off.

[37] Hans Herzberger, "Notes on Naive Semantics", and Anil Gupta, "Truth and paradox", both in *Journal of Philosophical Logic* 1982; and both reprinted in R. L. Martin, 1984. In what follows, I assume familiarity with these theories.

[38] Gupta is quite explicit about this: "We have used it [the notion of 'stable truth'] in the metalanguage to give an account of the concept of truth in the object language L" (*op. cit.*, p. 233).

[39] See, for example, Gupta's comments on p. 225: ". . . it is reasonable to believe that [the stable sentences] include all the unproblematic sentences", and "[t]he problematic sentences — such as the Liar and the truth teller ("This very sentence is true") — are, by our account, unstable".

[40] Gupta, *op. cit.*, p. 233. This is all Gupta says on the matter.

[41] And there are analogous difficulties. Many have found Tarskian accounts of 'true' artificial because the truth predicate is split into infinitely many distinct predicates; as Kripke puts the point: "Surely our language contains just one word 'true', not a sequence of distinct phrases $\ulcorner \text{true}_n \urcorner$, applying to sentences of higher and higher levels" (Kripke, *op. cit.* p. 58). There is an analogous problem for Gupta's extended theory: we can argue that in natural language there is just one predicate 'semantically unproblematic truth', and it is artificial to split it into infinitely many distinct predicates, each defined with respect to a distinct language. Further, how are we to connect these distinct predicates to ordinary uses of 'semantically unproblematic truth'? Which of these predicates is appropriate for the interpretation of, say, "No sentence is both a semantically unproblematic truth and a semantically unproblematic falsehood"?

[42] Nicholas Rescher and Robert Brandom, *The Logic of Inconsistency*, APQ Library of Philosophy, 1979.

[43] See *The Logic of Inconsistency*, pp. 34–35.

[44] See *The Logic of Inconsistency*, Section 26 and p. 4.

[45] It would seem that a *genuine* inconsistency solution to the Liar must dispense with the object language/metalanguage distinction altogether. Graham Priest's "paraconsistent" approach claims to do just this. This, I would argue, leads to new difficulties; but for reasons of space I shall not pursue this matter here.

[46] For reasons of space, I have not considered Feferman's classical modification of Kripke's theory (Solomon Feferman, "Towards Useful Type-Free Theories, I", *Journal of Symbolic Logic*, 1982, and in Martin 1984). As in Kripke's theory, the complement of truth is not expressible in the object language, and so the answer to the first of our questions is affirmative. And since Feferman's account does not prevent the formation of the counterdiagonal concepts, any more than Kripke's does, the answer to our second question is affirmative too. I cannot pursue the details here.

[47] Apart from Herzberger, discussed below, Donald Davidson also takes this view, in "Truth and Meaning", *Synthese* 1967.

[48] See Herzberger, "Paradoxes of Grounding in Semantics", and "New Paradoxes for Old", *Proceedings of the Aristotelian Society* 1981.

[49] Herzberger writes: "Formally, an elementary conceptual system is a triple $\{T, N, F\}$ consisting of a denumerable set T, its terminology; an abstract class N, its ontology; and a function f whose value for each term in T is a subset of N. A concept is expressible just in case the semantic rules f assign its extension to at least one of the terms of the system" ("Paradoxes of Grounding in Semantics", p. 157). This characterization of expressibility (and inexpressibility) carries over to the variously enriched conceptual systems which Herzberger goes on to develop.

[50] See "New Paradoxes for Old", pp. 113–114. As Herzberger points out, this negative expressibility claim is "relative to a very simple view of concepts and the way they

relate to words" (p. 113). Below, I shall argue that this claim cannot be made out, even relative to this simple view.

[51] See "Paradoxes of Grounding in Semantics", p. 153, and pp. 159–160. Herzberger defines a grounded term as a term that heads no infinite sequence of terms each of which is satisfied by its successor in that sequence.

[52] R. L. Martin, in "Are Natural Languages Universal?", has also argued against Herzberger's inexpressibility claim, though not along the lines I suggest below.

[53] The argument I have presented carries over in an obvious way to the case of grounded terms, and to the other inexpressibility results Herzberger obtains in "New Paradoxes for Old".

[54] I would extend this criticism to Tarskian theories, though this is not the place to develop the point. In a word, the stratification of, say, English into a hierarchy of distinct languages seems artificial, an overly rationalized account of natural language.

[55] Elsewhere, I propose a solution to the Liar which, I argue, satisfies both these criteria (see my "On a medieval solution to the Liar paradox", *History and Philosophy of Logic*, Volume 8, 1987, Number 2, and "A Singularity Solution to the Liar", in preparation).

[56] I would like to thank Tyler Burge, Bill Hart, David Kaplan, Gary Mar, Tony Martin, and an anonymous referee.

Department of Philosophy,
Caldwell Hall,
University of North Carolina,
Chapel Hill, NC 27599-3125,
U.S.A.