

Preface

I trust that I am not absolutely the last human being on the face of the Earth who might be expected to write a book on truth and the Liar paradox, but my name would not be among the first dozen, or first score, or first hundred to come to mind as a likely candidate for such an undertaking. I am not trained as a logician, or a specialist in semantics. Some explanation is in order.

I did not set out to write a book on the Liar paradox. Indeed, I did not set out to write a book at all, and the non-book I did set out to write was not about the Liar paradox. So what you have before you is the end product of a very long and difficult struggle in which one problem led to another and yet another, with the problem of truth ultimately emerging as the center of attention. The original project was concerned instead with the attempts by John Lucas and, more recently, Roger Penrose to draw consequences from Gödel's incompleteness proof about the structure of the human mind. The basic line of argument is well known: suppose that human reasoning capacity can be reduced to some sort of algorithmic procedure, such that the sentences about, say, arithmetic that one takes to be provably true with certainty can be characterized as a recursively enumerable set of sentences. Gödel's procedure then shows how to identify a sentence which is certain to be both true and not identified by the system as a truth if the system is consistent. The idea is now this: We can recognize the Gödel sentence of the system as true even though the system itself cannot. Therefore *our* insight into what is certainly true outruns that of the system. But the system has not been characterized in any way except to say that it is, in some sense, algorithmic. Therefore our insight is, in principle, more extensive than that of any algorithmic system. Therefore

our insight cannot be the result of any (consistent) algorithmic system. Therefore the power of our minds cannot be captured by any algorithm. Therefore we are not like computers. Yet further, according to Penrose, it follows that the physics which governs our brains cannot even be computable, otherwise our insights would, in the relevant sense, be the output of an algorithm (viz. the algorithm which specifies the dynamics of the physics of the brain).

There are obviously lots of lacunae in this argument, and the foregoing sketch is only the barest skeleton of the complete defense of the conclusion. But even without going into the details, there is something extremely odd about the conclusion. It seems to rely on the idea that our insight into the truth of the Gödel sentence of a (consistent) system is evidence of some almost mystical intellectual power, a power which cannot even be mimicked by a computer. But when one looks at the reasoning which leads to the conclusion that the Gödel sentence is true, one finds nothing remarkable or difficult or unformalizable. What one finds instead is a simple *reductio ad absurdum*. One constructs a sentence which says, in effect, that it is not a theorem, or recognized as an unassailable truth, by the algorithmic system in question. Then one asks whether the sentence in question is in fact recognized by the system as an unassailable truth. If it is, then the system in question judges as an unassailable truth a sentence which is, in fact, quite obviously false. So the system is not consistent, or at least makes obvious mistakes in what it recognizes as certainly true. Such a system cannot represent our reasoning power, since we would not make such an obvious mistake. So any consistent system cannot judge the sentence in question to be obviously true. But then, as we can see, the sentence *is* obviously true. So, providing the system is consistent, we can grasp more truths than it can.

Let us try to formulate the *reductio* in a rigorous way. We need a predicate $P(x)$ which will stand for "x is recognized as certainly true by the system" (if the system can be reduced to an algorithm, $P(x)$ will ultimately specify the algorithm, but this will do for now). Then we can characterize the Gödel sentence of the system, which we will call \square as $\square P(\square)$. The sentence \square says of itself that it is not provable by the system. Now suppose that we accept that the system in question is absolutely reliable: everything it recognizes as certainly true is, in fact, true. Then we will endorse the inference from the fact that a sentence is provable by the system, or recognized as true by the system, to that sentence itself. That is we will, in general, accept as valid the inference from $P(n)$ to the sentence denoted by n (I am being cavalier about the use/mention distinction here, but in an obvious way). Let us call that inference rule *Trust*, reflecting the fact that we ourselves trust every sentence the system can prove. (If we don't trust the system, e.g. if we think the system might be inconsistent, then the whole argument breaks down). To employ the inference rule, we need an algorithm which can determine what sentence is denoted by an individual term like \square , but in this case it is easy to specify this: \square denotes the sentence $\square P(\square)$.

Now we can reproduce the *reductio* we use to establish the truth of the Gödel sentence by a simple four line proof in a natural deduction system:

$P(\square)$	Hypothesis
$P(\square)$	Reiteration
$\sim P(\square)$	Trust
$\sim P(\square)$	~ Introduction

Since $\neg P(\ulcorner \neg P(\ulcorner) \urcorner)$ is the sentence denoted by $\ulcorner \neg P(\ulcorner) \urcorner$, we have managed to prove $\neg P(\ulcorner \neg P(\ulcorner) \urcorner)$ and simultaneously to prove that the system cannot prove $\neg P(\ulcorner \neg P(\ulcorner) \urcorner)$, since that is just what the sentence derived says. So we can prove more than the system.

There is nothing mystical or non-algorithmic about the reasoning just reproduced: it requires nothing more than the syntactically specifiable inference rules Trust and $\ulcorner \urcorner$ Introduction. (We count Trust as syntactically specifiable since we are allowing information about the denotation of $\ulcorner \urcorner$ into the inference rules themselves.) And similar reasoning can be produced for any predicate for which we accept the analog to the rule Trust, that is, for any predicate such that we accept that every sentence which satisfies the predicate is true. If we believe that, then we believe that the rule Trust is valid. In other words, the proof above seems to give us a general recipe: find any property of sentences such that every sentence which has that property is true, then we can construct a sentence which *does not* have that property but which we can recognize (by an analog of the proof above) as true. The *reductio* seems to have a magical power to reveal the truth of a sentence which does *not* have whatever truth-guaranteeing property we can specify: namely the sentence which says of itself that it does not have the truth-guaranteeing property.

Now this is just too good to be true. The argument is too general, and too simple, to really have the power ascribed to it. It occurs to one, in the first place, that we might as well let $P(x)$ stand for "sentence which *I* can recognize, by whatever means, as unassailably true". Now the little argument seems to prove, beyond any doubt, that $\ulcorner \neg P(\ulcorner \neg P(\ulcorner) \urcorner) \urcorner$ is true and (hence) that $\ulcorner \neg P(\ulcorner \neg P(\ulcorner) \urcorner) \urcorner$ cannot be recognized as unassailably true. For I must surely endorse the principle Trust in this case: it says that if I recognize a sentence as unassailably true then it is true. If I do not endorse Trust, then I recognize that there are sentences which I regard as

certainly true but which I suspect may not be true: a clear contradiction. So the little proof above would seem to prove, to my own satisfaction, the truth of \square . Hence, I can apparently prove (to my own satisfaction) a sentence which, if true, I cannot prove to my own satisfaction. But things get worse.

We have said that the little proof goes through for any predicate that denotes a property of a sentence which guarantees truth, and that we must regard the proof as going through for any such predicate which we regard as denoting such a property (so we accept the rule Trust). But the most obvious and impermissible such predicate is "true". If we accept the argument above, letting $P(x)$ stand for "true", then we seem to be able to prove the truth of a sentence to which that predicate does not apply, i.e. we seem to be able to prove a sentence which (if our proofs are to be trusted) *is not true*. Further, if we let $P(x)$ stand for "true", then \square is a sentence which says *that it is not true*, i.e. it is a classical Liar sentence. Our original concern with Gödel's argument has led us back to the Liar.

This route to the Liar is rather different from the usual approaches. One typically begins by considering the Liar sentence, and arguing that one cannot consistently hold it to be true (since that leads to a contradiction) and cannot consistently hold it not to be true (since that leads to a contradiction as well). That is, indeed, a puzzle. But what is even more puzzling, and more upsetting, is that we apparently can *prove* both the Liar and its negation, so if we really trust the standard system of inferences, we are forced to accept both the Liar and its negation. This is what will be called the *Inferential Version* of Liar paradox. What the Inferential Version really shows, of course, is that we cannot consistently accept the validity of all of the inferences which we intuitively take to be valid. We must somehow amend or restrict logic to keep ourselves from falling into contradiction.

My original suspicion was that the very modifications to our inference schemes needed to save us from the inferential paradox might also undercut the arguments which we use to convince ourselves of the truth of the Gödel sentence of a formal system. The informal Gödel reasoning is so close to our reasoning about the Liar that it seemed likely that whatever error infects the latter also infects the former. My plan was to fix up logic to escape from the Liar problem, and then see how the necessary modifications affect arguments which depend of Gödel's results. The project did not proceed smoothly.

My first attempt to solve the inferential problem still resides on my hard disk, in a file entitled "Truth". It comes to an abrupt end after 78 pages. The original scheme, of which nothing now remains, never exactly died, but it became progressively more cumbersome and baroque. Every solution to one problem engendered another difficulty, which could in turn be addressed. Conditions piled on conditions and clauses on clauses, with no end in sight. Eventually, the attempt was simply abandoned. I salvaged what I could, and started a second file, entitled "New Truth".

"New Truth" pattered along for 124 pages before finally giving out. I had hit on the solution to the inferential problem, but was only slowly coming to realize that defending the solution required one to tackle the problem of truth directly, by producing a semantics for the language. The problem is fairly simple: in defending a set of inference rules, one would like to show that they are truth-preserving. But without an explicit theory of truth, this cannot be done. So "New Truth" had to be retooled.

The next file was entitled "Final Truth". It soldiered on for 164 pages before being abandoned. "Final Truth" contains the key to the semantic theory which ultimately survived: the idea of the graph of a language, and the analogy between the problem of semantics and boundary value problems as

they appear in mathematics and mathematical physics. But the means for avoiding the Liar paradox were still clumsy and complicated, and the restrictions imposed on the language felt *ad hoc*. "Final Truth", I decided, was too much of a contraption to be the final truth.

Running out of both patience and superlatives, I opened a file called "Ultimate Truth". That file contains the text before you. The theory in "Ultimate Truth" is vastly more compact and elegant than any of its forebears, and I had hoped that the end result would be a more compact paper. But new ideas and applications kept crowding in, new paradoxes came forward and were resolved, the scope of the paper enlarged. In the end, Gödel's theorem, the original object of inquiry, takes up a scant few pages. But having gone through so many different ideas, and abandoned so many different approaches, I feel secure that there is some value to what survives.

Composing a book by fits and starts, wandering into cul-de-sacs and then starting over again, is not an intrinsically pleasant enterprise. Still, writing this book has been as enjoyable as such a tortuous journey could be. The bulk of the text was written in the first half of 1997, while on sabbatical in Cambridge Massachusetts. Following a tradition set the previous semester when I had been visiting at Harvard, Ned Hall, Jim Pryor, Jennifer Noonan and Justin Brookes would join my wife Vishnya, my daughter Clio, and me weekly for a banquet of gourmet food (Vishnya's masterpieces) and conversation. Any mild disaster becomes comic in the telling. In such delightful company, the seemingly endless series of failures, reversals and dead-ends on my path to this theory was transformed from a chronic frustration to a serialized shaggy dog story. Ned especially was taxed with every gory detail, and his boundless

acuity and cheer made the whole process not merely bearable, but fun. His insights were always sharp and to the point. Vann McGee provided very valuable comments on the penultimate draft, and pushed me to face the Permissibility Paradox directly, rather than trying to finesse it. It was also a great help to be able to present some preliminary results at the London School of Economics. And without the generous support of Rutgers University during the sabbatical, this work would never have been done.

The final draft has benefited greatly from the generous comments of Harty Field, and the joint comments of a reading group at MIT and Harvard. I am also very grateful to the Department of Philosophy at University of Massachusetts, Amherst for allowing me to present an overview of the final theory, and vigorously challenging the approach.

My greatest, unrepayable debt is to Vishnya, who lived through every modification, retraction, and short-lived enthusiasm, day after day, for months. I would be fortunate beyond measure if only for her constant affection, but to have in addition a partner in life with whom a project like this can be shared is more than anyone could deserve.