Etica & Politica / Ethics & Politics, 2003, 1
http://www.units.it/~dipfilo/etica_e_politica/2003_1/4_monographica.htm

Satan stultified: a rejoinder to Paul Benacerraf (*)

J.R. Lucas

Fellow of Merton College, Oxford
Fellow of the British Academy

The application of Gödel's theorem to the problem of minds and machines is difficult. Paul Benacerraf (1) makes the entirely valid `Duhemian' point that the argument is not, and cannot be, a purely mathematical one, but needs some philosophical premisses to be able to yield any philosophical conclusions. Moreover, the philosophical premisses are of very different kinds. Some are concerned with what is essential to being a machine---these are typically intricate, but definite, easily formalised by the mathematician, but unintelligible to the layman: others attempt to capture what is essential to being a mind, a person or a self---these are typically intuitive, but vague; resistant to exact definition by the logician, but, none the less, widely used and well understood. Gödel's theorem itself, like many other truths, can be taken either way: it can be taken as a formal proof sequence yielding certain syntactical results about a certain class of formal systems, but it can also be taken as giving us a certain type or style of argument, which we can understand, and, once having got the hang of it, adapt and apply in innumerable different circumstances. In my dispute with the mechanist, I take Gödel's argument both ways: I first take it as an argument which the mechanist, even according to his own mechanist principles, must accept as scoring some point against his favourite machine; and then I hope that the mechanist, as a man, will see that he can do better and that this sort of argument will always apply against any form of mechanism he espouses.

The argument is a dialectical one. It is not a direct proof that the mind is something more than a machine, but a schema of disproof for any particular version of mechanism that may be put forward. If the mechanist maintains any specific thesis, I show that [146] a contradiction ensues. But only if. It depends on the mechanist making the first move and putting forward his claim for inspection. I do not think Benacerraf has quite taken the point. He criticizes me both for "failing to notice" that my ability to show that the Gödel sentence of a formal system is true "depends very much on how he is given that system" (2) and for putting the argument in the form of a challenge in which I challenge the mechanist to produce a definite specification of the Turing machine that he claims I am. (3) Benacerraf thinks that the argument by challenge reduces the argument to a mere contest of wits between me and the mechanist. But we are not trying to see who can construct the smartest machine, we are attempting to decide the mechanist's claim that I am a machine: and however clever the mechanist is, even if he were not a mere man but Satan himself, I, or at least an idealised and immortal I, could out-Gödel it, and see to be true something it could not. Benacerraf protests that "It is conceivable that another machine could do that as well." Of course. But that other machine was not the machine that the mechanist was claiming that I was. It is the machine which I am alleged to be that is relevant: and since I can do something that it cannot, I cannot be it. Of course it is still open for the mechanist to alter his claim and say, now, that I am that other machine which, like me, could do what the first machine could not. Only, if he says that, then I shall ask him "Which other machine?" and as soon as he has specified it, proceed to find something else which that machine cannot do and I can. I can take on all comers, provided only they come one by one in the sense of each being individually specified as being the one that it is: and therefore I can claim to have tilted at and laid low all logically possible machines. An idealised person, or

mind, may not be able to do more than all logically possible machines can, between them, do: but for each logically possible machine there is something which he can do and it cannot; and therefore he cannot be the same as any logically possible machine.

Benacerraf attempts to reconstruct the argument not as a dialectical one but as a simple proof-sequence using only formally defined [147] terms. It is helpful and illuminating to follow this approach through: but it is a distortion of the original argument, and many of Benacerraf's criticisms are criticisms not of arguments I actually put forward, but of ones he from a very different point of view thinks I should have put forward. He complains that I equivocate and never make clear the sense of the word `prove' in which I claim to be able to prove things which a machine cannot prove. (4) And indeed I never do elucidate any such absolute sense of the word `prove'---for I took care not to use the word in such a sense, but to frame the contrast between what was provable-in-the-system (the machine's system) and what could be produced by me as true. This is not to say, of course, that Benacerraf's locution is improper or unintelligible. Many people do speak colloquially of their being able to prove things a machine cannot prove, and it would be permissible to define an absolute provability to accommodate this locution. Permissible, but inexpedient. Provability has been construed by mathematical logicians for a generation as a syntactical term with a very precise definition, and it could be confusing to import loose notions of what I can see to be true into this well-disciplined and useful concept.

Benacerraf does not see why the limitation on what a machine (which he calls Maud) can do should matter. "Maud is limited in the things for which she can offer $\text{formal}_{\text{Maud}}$ proofs. But does this Limit the *informal* (i.e. not formalizable in Maud) proofs she can conjure up?" (5) But this is to forget that Maud is a machine. Machines cannot conjure up anything, but only act according to their input and the way they are wired up. Mechanists maintain that the same is true of men, and that all a man's intellectual output is, in fact, the output of some machine, and so is simply part of some *formal* system. Although a machine might be programmed *both* to produce formal proofs in some specified formal system *and* to go through various logical manoeuvres which would resemble those of a woman convincing herself of things, *all* the operations of the machine, of either sort, would be governed by the programme, and would correspond to a formal system, though not necessarily the one specified in the programme for `proving' manoeuvres. For a [148] machine, and on the mechanist thesis for men too, `producing as true' is simply and entirely the consequence of certain antecedent conditions, and therefore can be represented as being proved-in-a-formal-system of some sort. Some people will impose a definitional stop at this point, and say that, by definition, only a mind or a person can be said to produce anything as *true*, but such a move will cut no ice with a mechanist. In order to argue with him we must go along with him to the extent of allowing that we could talk of a machine (either human or artificial) `producing something as true', but that this would have to be construed on his thesis as being proved-in-a-formal-system of some sort, and that therefore anything a machine could produce as true would have to be something that was provable-in-the-system. The mechanist, regarding men as something less than men, namely machines, regards their concept of truth as again something less, namely provability-in-a-given-system. But it is implausible to reconstruct truth as provabilityin-a-given-system, and Gödel's theorem shows this (or, at least, that the philosophical price of identifying truth with provability-in-a-system would be very high). "If truth is not provability-in-a-given-system" a tough-minded philosopher may ask "what is it?." I cannot answer him. I think I know what truth is, but I know I cannot tell anybody else exactly what it is. I can recognize---subject to many mistakes, errors and oversights---various propositions, formulae, statements and theories as true, or come to that conclusion after due consideration. In particular cases I can explain why, and often hope to convince somebody else as well. But I cannot produce a formula which

will cover all cases, or frame an instruction which somebody else could apply mechanically in all cases---indeed, as Tarski has shown, any attempt to give a formal representation of truth must lead to a contradiction. (6) I can communicate the idea to him, but he will need, on occasion, to use his own *nous* and make up his mind [149] for himself. If it were not so, I doubt if we could regard anybody else as being *anybody* else; a listener I could completely instruct in all the truth would not seem to have a mind (be any*body* rather than any*thing*) independently of me (be else, or *other than* me). Some degree of intellectual autonomy is a necessary condition of being a rational agent. And therefore truth cannot be precisely defined.

Instead of pressing me with the question `What is truth?', Benacerraf seeks to reconstruct my argument as formally as possible, in order to determine exactly what philosophical assumptions are involved and might be dispensed with. Gödel's first theorem, taken entirely formally, shows only that in any formal system rich enough for elementary arithmetic a formula can be found which cannot be proved-in-that-system, unless the system is inconsistent, and cannot be disproved-in-that-system, unless the system is omega-inconsistent; that is, it shows only an inconsistency between the syntactical properties of completeness and omega-consistency, not anything to show a difference between the semantic concept of truth and the syntactical one of provability. Benacerraf is therefore precluded from using Gödel's first theorem, and turns instead to its corollary, Gödel's second theorem, that in any formal system rich enough for elementary arithmetic, the consistency of that system cannot be proved-in-that-system, unless the system is in fact inconsistent. If the mechanist thesis implies that a machine that is equivalent to me must be able to prove its own consistency because I can establish mine, then it would yield a contradiction, and some part, at least, of the mechanist thesis would have to be given up.

But it might not have to be the essential part. Benacerraf distinguishes three components, 9(a), 9(b) and 9(c), (7) and argues that instead of abandoning the essential component 9(c), that for each human being there is some definite Turing machine, W_j, which can prove-in-its-formal-system everything that that particular human being can produce as true, we might jettison the second half of 9(b) or of 9(a), and say only that I cannot prove the consistency of the machine, or, less plausibly, that I cannot prove that the machine is adequate for arithmetic.[150] In either case Benacerraf's strategy is to deny me a premiss I need, not by *denying* that the premiss is true, but simply by *not conceding* that it is. My argument depends on knowing what sort of Turing machine I am alleged to be: "It will not suffice," he says, "that W_j merely *be* a subset of my output, *I must be able to identify it as such and by that name*" (8) for a contradiction to ensue. "If given a black box and told not to peek inside, then what reason is there to suppose that Lucas or I can determine its program by watching its output? But I must be able to determine its program (if this makes sense) if I am to carry out Gödel's argument in connection with it" (9) and therefore Benacerraf counters with the suggestion that I am, indeed, a Turing machine, but I know not which. He hopes thus to keep open an escape route from the contradiction the mechanist thesis leads to, and instead of having to abandon the essential 9(c), deny me the second half of 9(b) or 9(a); so that not knowing what sort of Turing machine I was alleged to be, I could not prove its consistency or its adequacy for arithmetic, even though in fact it was both consistent and adequate.

But it will not do. Benacerraf's mechanist seeks to avoid defeat by playing with his cards very close to his chest, in order not to concede the premiss I need, but has to hold them so close that he does not play them at all. In response to the Delphic injunction, *gnothi seauton*, KNOW THYSELF, he returns an answer so deviously Delphic as not to say anything at all. "You are a machine" he suggests, but when one asks what sort of a machine, he replies that it is oneself. (10) But this is to evacuate the mechanist thesis of all content. I did not need the mechanist to tell me that I am I; nor need I be much put out by mechanism if that is all that it can say about

me. Only if, at least in principle, my programme could be known and all my actions infallibly predicted and mechanistically explained, can mechanism worry us. Benacerraf's mechanism, which avoids contradiction only by never specifying what sort of machine a human being is alleged to be, is, from a common sense point of view, a position too empty to be worth holding. In fact it is also logically untenable. In his efforts to deny me my [151] dialectic, Benacerraf's mechanist is self-stultifyingly eristic, and is guilty of omega-inconsistency. He maintains that there is a programme number **j** such that the corresponding programme, W_j, represents me, but must he careful not to particularise and specify what sort of machine I am. For if he says which W_j it is that satisfies his conditions 9(a)(i)Q *is included in* W_j, 9(b)(i) W_j *is included in* S*, and 9(c), then I can check up on his claims 9(a)(i) and 9(b)(i), and see whether or not they are true; and if they are, then I shall know that they are, and so `Q *is included in* $W_j$' *is a member of* S and `W_j *is included in* S*' *is a member of* S, and a fortiori `Q *is included in* $W_j$' *is a member of* S* and `W_j *is included in* S*' *is a member of* S*: that is, the full conditions 9(a) and 9(b) will both hold together with 9(c) and a contradiction will ensue. But to maintain that there is a programme number **j** such that the corresponding programme W_j represents me, while knowing that for each particular programme number **j** there is an argument, different in each case, showing that that W_j does not represent me, is to be omega-inconsistent. Benacerraf is claiming that the man is a machine, although for every particular machine he could be we can show that he is not that one. It is not simply that, as Benacerraf suggests, "*I am indeed a Turing machine but one with such a complex machine table (program) that I cannot ascertain what it is*" (11) my ignorance is more necessary than that, and more fatal to his thesis. The only way I can be absolutely sure of not knowing that I am any particular machine is by not being any machine whatever.

In order to avoid omega-inconsistency Benacerraf needs to shift his ground, and prevent the anti-mechanist from knowing, not what sort of machine he is alleged to be, but whether it is consistent or not. For Gödel's theorem applies only to consistent systems, and if we are to apply it to a given system we must be able to say that it is consistent. In some cases we may be able to do this directly. Gentzen proved the consistency of elementary number theory by transfinite induction: and although the mechanist will counter by saying that a machine could do as much, it would have to be a different machine from the one originally under discussion, and the one originally under discussion would have been shown not to be equivalent to me. However, although we always may be able to produce a direct consistency proof, we have no general recipe by means of [152] which we can always be sure of producing such a proof. And therefore we have to turn to indirect arguments.

One way of checking the consistency of a system is through the Gödelian formula itself. The mechanist---let us grant to Benacerraf that he is no mere man, but the Prince of Darkness himself (12) ---produces the specification of a machine which he claims is equivalent to me. From the specification, I calculate the Gödelian formula, and I ask the mechanist (not the machine) whether it is one the machine can prove-in-its-system. The mechanist should know. After all, he claims to know the machine well enough to know that it is equivalent to me. If he cannot answer even this single question, his claim will look somewhat suspect. Yet it appears nevertheless that the mechanist, even if Satan himself, is necessarily ignorant. For if he knew, whichever answer he gave would invalidate his claim. If he said that the machine could prove-in-its-system the Gödelian formula, then I should know that the system was inconsistent, and so could not be equivalent to me: while if he said that the machine could not prove-in-its-system the Gödelian formula, then I should know, since there was at least one well-formed formula it could not prove-in-its-system, that it was absolutely consistent and so that the Gödelian formula was true, although the machine could not prove-it-in-its-system, and hence that the machine was again not equivalent to me. A mathematical theologian who was also a mechanist might derive the comforting conclusion that either the Devil does not exist or at least he is not omniscient.(13) But it would not save mechanism. So long as there is, or could be,

any person in the Universe---even myself---who does, or might, maintain that I am equivalent to a machine, I can use a Gödelian question to reveal whether the machine is consistent or not, and, either way, to show mechanism to be false.

It is necessary to ask the mechanist, not the machine. The machine cannot answer the question whether it can prove---, or cannot prove---, the Gödelian formula in-its-system. But the question is an askable one, and one which we can press on the mechanist, who, [153] being a person, whether human or Satanic, cannot convincingly plead ignorance. For ignorance is of no avail when impaled on the horns of a dilemma. Either the machine can prove-in-its-system the Gödelian formula or not: if it can, it is inconsistent, and not equivalent to me; if it cannot, then I can produce the Gödelian formula as true, and am not equivalent to the machine. Either way, I and the machine are not equivalent to each other, and the mechanist thesis fails.

The same dilemma can be posed simply in terms of consistency. Every formal system and every Turing machine is either consistent or inconsistent. If the latter, it is degenerate, not a proper system or machine at all, and not a plausible candidate for being any sort of model of my mind: only if it is consistent, and acknowledged as such, is it even a candidate, and then, being consistent, and being said to be consistent, I can out-Gödel it. Benacerraf and Putnam (14) seek to evade this dilemma by representing it in the form of a conditional in which the antecedent can never be asserted as an independent premiss in the form required. Gödel's first theorem can be expressed by saying that for any formal system **T** (Putnam's version) or Maud (Benacerraf's version), we can prove (in a strict, syntactical sense) :

If **T** is consistent, **U** is true.

But this proof, being a formal one, can be represented in **T**, and, as in Gödel's second theorem, we can prove-in-**T** the implication

If **T** is consistent, **U** is true,

so that the difference between **T** and me seems to have disappeared. Only if I can detach the antecedent and assert that **T** is consistent, can I assert categorically that **U** is true, and do something that cannot be done in **T**, and so show myself different from **T**. How do I know that **T** is consistent? After all, **T** may be very complicated---if it is to represent me, it must be---and even though we have been able to find consistency proofs for some formal [154] systems, there are many others we have not as yet been able to prove consistent. If **T** were equivalent to Quine's New Foundations, could I prove its consistency? Whence, then, do I obtain the premiss that **T** is consistent? I answer "From the mechanist himself." It is his claim which provides me with the premiss, for it is a claim that **T** is equivalent to me, and unless **T** were consistent it could not be. If the mechanist advances the claim that a particular machine, represented by a formal system **T**, is equivalent to me, then I can say "So you say **T** is consistent?" and he must concede it, or else abandon his claim that **T** is equivalent to me. But once he has granted that **T** is consistent, I can produce the Gödelian formula of **T** as true; which **T**, if it is consistent, cannot do, and hence I am not equivalent to **T**.

The argument is dialectical. It is an argument between two persons, not a proof sequence constructed by one. I believe that the dialectical form reveals the underlying logic better than any monologue can, both here and earlier, when we needed to explicate the different meanings of the word `a'. But logic has been traditionally monologous, (15) and we need to see how this argument would appear in monologous form. Putnam suggested (16) that it consisted of the five steps

(1) If **T** is inconsistent, then not-(**T**=me)
(2) See_I (If **T** is consistent, then **U**)
(3) If **T** is consistent, then See_I (**U**)
(4) If **T** is consistent, then not-See_T (**U**)
(5) If **T** is consistent, then not-(**T**=me)

where See_I(...) means `I can see it to be true that ... ' and $See_T$(... ) means `**T** can produce it as being true that. . .' Putnam objects that the third proposition does not follow from the second,

and that therefore the whole proof-sequence is invalid; and since I have given no formal rules for the use of `I can see it to be true that . . .', and indeed, have said that such a notion cannot be completely formalised, (17) his attack is strong. [155]

The monologous form of the dialectical argument I have been deploying is *reductio ad absurdum*. In default of an actual mechanist to argue with, I myself entertain the hypothesis that mechanism is true, and following out the consequences find they lead to contradiction. If I supposed mechanism were true, I should have to suppose that there was some particular system **T** which represented me, and was therefore consistent. This being so, I should conclude that the Gödelian formula **U** was true, although I should also be led to conclude that **U** was unprovable-in-**T**. These conclusions, although only hypothetically entertained, would make mechanism a sufficiently uncomfortable position for me to abandon the idea of taking it up. In Putnam`s formulation we need to place an `I can see it to be true that . . .' in front of propositions 1, 3, 4, 5. We then have

(1') See_I (If **T** is inconsistent, not-(**T** = me))
(2) See_I (If **T** is consistent, then **U**
(3') See_I (If **T** is consistent, then See_I (**U**)
(4') See_I (If **T** is consistent, then not-(See_T (**U**))
(5') See_I (If **T** is consistent, then not-(**T** = me)).

We then have to ask whether if, thanks to Gödel's theorem I can see that If **T** is consistent, **U** is true, I can further see that If **T** is consistent, I can see that **U** is true. And the answer is, surely, that I can, because I can *understand* Gödel's theorem, and see therefore that granted its premisses, I can see its conclusion to be true. I am not simply given the conclusion of Gödel's Theorem:

If **T** is consistent, then **U**

but am able myself to produce an argument showing **U**, provided that **T** is consistent. Reviewing this argument, I can see that granted **T** is consistent, I can follow Gödel's argument through and see that **U** is true. The step from (2') to (3') is therefore justified, and (5') follows. If I were a mechanist, I should have to accept both (1') and (5'), which together constitute a dilemma leading to *reductio ad absurdum*.

We can look at the step from (2) to (3) another way. I maintain that I should find it embarrassing to have to believe both proposition (2) and the negation of proposition (3). The negation of

(3) If **T** is consistent, then See_I (**U**)

[156] is

(3) Although **T** is consistent, not-(See_I (**U**)

that is, Although **T** is consistent, I cannot see that **U** is true. Putnam's argument is that although **T** is consistent, I may not know that it is, and therefore cannot see that **U** is true, although I can see that if **T** be consistent, **U** is true. Putnam's argument is fair, put from the third-personal point of view. But moods and persons are closely interlocked here. Putnam can say of Lucas, that Although **T** is consistent, Lucas (who is not entitled to assume that it is) cannot see that **U** is true: but if I want to do the saying myself, I have to say that Although **T** *may be* consistent, I (who am not entitled to assume that it is) cannot see that **U** is true. In changing the third to the first person, I have also to change the categorical indicative to a hypothetical subjunctive. Else the concession made in the `although' clause belies the rider in brackets, which is essential for Putnam's conclusion: if I am not entitled to assume that **T** is consistent, I must not assume that it is---though I may assume that it may be. I can believe

(~~3~~) Although **T** may be consistent, not-(See_I (**U**)) as well as

(2) See_I (If **T** is consistent, then **U**), but I cannot myself believe both (~~3~~) Although **T** is consistent, not-(See_I (**U**)) and (2) See_I (If **T** is consistent, then **U**), for *if* I am making. the categorical assumption expressed by the clause `Although **T** is consistent', *then* I *can* see that **U** is true, in virtue of what I can see in (2). And this is what I am doing. I am considering the

possibility that I am a machine represented by a formal system **T**. There are then two alternatives, that **T** is inconsistent or that it is consistent. I assume each in turn. Even if I do not know which assumption actually holds, I am entitled to make each assumption in turn, and follow out that assumption to its conclusions, working out what can be established, granted that assumption. In this procedure, while I am making an assumption, I am entitled to assume it. Hence I should be contradicting myself, if, while I was exploring the consequences that followed from the assumption that **T** was consistent, I had to deny that I could see the truth of the Gödelian formula on the grounds that **T** might not be consistent after all and in any case I was not entitled to assume that it was. Thus, although a third person talking about me could affirm (2) and deny (3), I cannot; [157] nor could he, if the `I' of (2) referred to him. So neither I nor he can accept mechanism without being involved in absurdity. So mechanism must be false.

The only line of escape left to the mechanist is to maintain that we are, or at least may be, inconsistent machines, or, even if we are not, cannot say that we are not. In my original paper (18) I argued against the viability of such a thesis, but Benacerraf finds my arguments unconvincing. (19) Essentially, I argue that human beings, and rational agents generally, are selective. I am not prepared to say just anything. I operate a distinction between those propositions I regard as true and those I regard as false, and am prepared to affirm the former but not the latter. In this I conform to Tarski's definition of absolute consistency, (20) and when in argument with you I explode "If you say that, you would say anything" I am avowing absolute consistency as a condition of rational discourse. But am I entitled to? Putnam and Benacerraf are sceptical, in the absence of a formal consistency proof. I may feel I am consistent but in fact be as inconsistent as naive set-theory was, or as Quine's New Foundations may be. But this is too static a comparison. When naive set theory was found to be inconsistent, it was replaced by Zermelo's. The version of New Foundations at present on the market is a reconditioned model: its predecessor was found to be inconsistent by Wang, and had to be returned for a manufacturer's overhaul. And this is typical. Inconsistency is the sort of infelicity up with which we will not put. By one device or another we obviate any inconsistency we come across. If any set of propositions we had hitherto accepted should turn out to be inconsistent, we should amend them so as to iron out that inconsistency, and we should continue to do this as often as need be. The intellectual output of a [158] rational agent is not one formal system, but, at least, a series of formal systems, generated by some principle of non-contradiction which, in view of the Gödelian argument, cannot itself be formalised.

It is rational for a rational agent to believe in his own rationality, and hence in his own consistency: both because it is the only assumption on which further thought is possible at all, and because it is not simply a matter of bare fact, but of decision. We decide to be consistent, to discipline our thinking and not to allow ourselves to affirm anything whatever, but to draw some distinction between truth and falsehood. I could not believe I was an inconsistent Turing machine without abandoning my conviction that there was a difference between truth and falsehood which I should always endeavour to follow in my own thinking, with some hope of occasional success. It is, in the language of mathematical theology, an act of faith. No formal proof is possible: but the alternative is self-stultifying. Although the possibility of my being essentially inconsistent cannot be ruled out by formal arguments and is perhaps just conceivable, it is a position of complete mathematical nihilism, which I, although neither a mathematician nor a theologian, am reasonably reluctant to adopt.

Notes

(*) First published: *The Monist*, vol.52, no.1, January 1968 pp.145-158. Republished here by permission. For the copyright of the papers of J.R. Lucas see http://users.ox.ac.uk/~jrlucas/ back

(1) Paul Benacerraf, God, The Devil, and Gödel, *The Monist*, 51 No. 1 (1967), 9-32. All simple page references refer to this article. back

(2) . P. 28. back

(3) Pp. 22-23. The reader should compare the quotation on p. 23 with the original. back

(4) PP. 19-21. back

(5) P.20. back

(6) See Alfred Tarski. *Logic, Semantics, Metamathematics*, trans. J. H. Woodger (Oxford, 1956), pp. 187-188, and 247; or Raymond M. Smullyan, *Theory of Formal Systems* (Princeton, 1961), chap. Ill, A , sec. 2, Theorem 1.1, p. 45. Compare Benacerraf's Appendix, pp. 30-32. Perhaps the moral we are intended to draw is that since the word `true' cannot be formally defined, it really has no sense and should be extruded from our vocabulary. But if the mechanist is by his own principles precluded from saying that mechanism is true, there is nothing he can say that could worry us. back

(7) P. 25. back

(8) P.25. back

(9) P.28. back

(10) P.28. back

(11) P. 29. back

(12) Pp.22-23 back

(13) Benacerraf himself (in conversation) professes agnosticism on this the intellectual powers and the ontological status of devils. I myself take the lowest possible view of both, but on other grounds. back

(14) Hilary Putnam, "Minds and Machines," in *Dimensions of Mind: a Symposium*, ed. Sidney Hook (New York: Collier edition, 1961), p. 142; reprinted in A. Ross Anderson (ed.), Minds and Machines (Englewood Cliffs, New Jersey, 1964), p. 77. back

(15) See J. R. Lucas, "Not `Therefore' but `But'," Philosophical Quarterly, 16 (1966),289-307. back

(16) At a discussion in M.I.T. on October 7th, 1967. back

(17) Pp. 147-148 of this essay. back

(18) J. R. Lucas, Minds, Machines and Gödel, *Philosophy*, 36 (1961), 120-124, reprinted in Kenneth M. Sayre and Frederick J. Crosson, eds., *The Modeling of Mind* (Notre Dame, 1963), pp. 263-268, and in Alan Ross Anderson, cd., *Minds and Machines* (Englewood Cliffs, 1964), pp. 52-56. back

(19) P.29 back

(20) A. Tarski, "Fundamentale Begriffe der Methodologie der deductiven Wissenschaften I" *Monatshefte fr Mathematik und Physik*, 37 (1930), 387-388; in Alfred Tarski, trans. J. H. Woodger, *Logic, Semantics and Metamathematics* (Oxford, 1956), p. 90. The most accessible account is Alonzo Church, *Introduction to Mathematical Logic* I (Princeton, 1956) chap. 1, sect. 17, pp. 108-109. back