Etica & Politica / Ethics & Politics, 2003, 1
http://www.units.it/~dipfilo/etica_e_politica/2003_1/12_monographica.htm

Minds, Machines and Gödel: a Retrospect (*)

J.R. Lucas

Fellow of Merton College, Oxford
Fellow of the British Academy

I must start with an *apologia*. My original paper, Minds, Machines and Gödel, was written in the wake of Turing's 1950 paper in *Mind*, and was intended to show that minds were not Turing machines. Why, then, didn't I couch the argument in terms of Turing's theorem, which is easyish to prove and applies directly to Turing machines, instead of Gödel's theorem, which is horrendously difficult to prove, and doesn't so naturally or obviously apply to machines? The reason was that Gödel's theorem gave me something more: it raises questions of truth which evidently bear on the nature of mind, whereas Turing's theorem does not; it shows not only that the Gödelian well-formed formula is unprovable-in-the-system, but that it is true. It shows something about reasoning, that it is not completely rule-bound, so that we, who are rational, can transcend the rules of any particular logistic system, and construe the Gödelian well-formed formula not just as a string of symbols but as a proposition which is true. Turing's theorem might well be applied to a computer which someone claimed to represent a human mind, but it is not so obvious that what the computer could not do, the mind could. But it is very obvious that we have a concept of truth. Even if, as was claimed in a previous paper, it is not the *summum bonum*, it is a *bonum*, and one it is characteristic of minds to value. A representation of the human mind which could take no account of truth would be inherently implausible. Turing's theorem, though making the same negative point as Gödel's theorem, that some things cannot be done by even idealised computers, does not make the further positive point that we, in as much as we are rational agents, can do that very thing that the computer cannot. I have however, sometimes wondered whether I could not construct a parallel argument based on Turing's theorem, and have toyed with the idea of a von Neumann machine. A von Neumann machine was a black box, inside which was housed John von Neumann. But although it was reasonable, on inductive grounds, to credit a von Neumann machine with the power of solving any problem in finite time---about the time taken to get from New York to Chicago by train---it did not have the same edge as Gödel's proof of his own First Incompleteness Theorem. I leave it therefore to members of this conference to consider further how Turing's theorem bears on mechanism, and whether a Turing machine could plausibly represent a mind, and return to the argument I actually put forward.

I argued that Gödel's theorem enabled us to devise a schema for refuting the various different mechanist theories of the mind that might be put forward. Gödel's theorem is a sophisticated form of the Cretan paradox posed by Epimenides. Gödel showed how we could represent any reasonable mathematical theory within itself. Whereas the original Cretan paradox, `This statement is untrue' can be brushed off on the grounds that it is viciously self-referential, and we do not know what the statement is, which is alleged to be untrue, until it has been made, and we cannot make it until we know what it is that is being alleged to be false, Gödel blocks that objection. But in order to do so, he needs not only to represent within his mathematical theory some means of *referring* to the statement, but also some means of expressing mathematically what we are saying about it. We cannot in fact do this with `true' or `untrue': could we do that, a direct inconsistency would ensue. What Gödel was able to do, however, was to express within his mathematical system the concept of being *provable-*, and hence also *unprovable-*, in-that-system. He produced a copper-bottomed well-formed formula which could be interpreted as saying `This well-formed formula is unprovable-in-this-system'. It follows that it must be both unprovable-in-the-system and none the less true. For if it were provable, and provided the system is a sound one in which only well-formed formulae expressing true propositions could be proved, then it would be true, and so what it says, namely that it is unprovable-in-the-system, would hold; so that it would be *un*provable-in-the-system. So it cannot be provable-in-the-system. But if it is unprovable-in-the-system, then what it claims to be the case is the case, and so it is true. So it is true but unprovable-in-the-system.

Gödel's theorem seemed to me to be not only a surprising result in mathematics, but to have a bearing on theories of the mind, and in particular on mechanism, which, as Professor Clark Glymour pointed out two days ago, is as much a background assumption of our age as classical materialism was towards the end of the last century in the form expressed by Tyndale. Mechanism claims that the workings of the mind can be entirely understood in terms of the working of a definite finite system operating according to definite deterministic laws. Enthusiasts for Artificial Intelligence are often mechanists, and are inclined to claim that in due course they will be able to simulate all forms of intelligent behaviour by means of a sufficiently complex computer garbed in sufficiently sophisticated software. But the operations of any such computer could be represented in terms of a formal logistic calculus with a definite finite number (though enormously large) of possible well-formed formulae and a definite finite number (though presumably smaller) of axioms and rules of inference. The Gödelian formula of such a system would be one that the computer, together with its software, would be unable to prove. We, however, could. So the claim that a computer could in principle simulate all our behaviour breaks down at this one, vital point.

The argument I put forward is a two-level one. I do not offer a simple knock-down proof that minds are inherently better than machines, but a schema for constructing a *dis*proof of any plausible mechanist thesis that might be proposed. The disproof depends on the particular mechanist thesis being maintained, and does not claim to show that the mind is uniformly better than the purported mechanist representation of it, but only that it is one respect better and therefore different. That is enough to refute that particular mechanist thesis. By itself, of course, it leaves all others unrefuted, and the mechanist free to put forward some variant thesis which the counter-argument I constructed does not immediately apply to. But I claim that it can be adjusted to meet the new variant. Having once got the hang of the Gödelian argument, the mind can adapt it appropriately to meet each and every variant claim that the mind is essentially some form of Turing machine. Essentially, therefore, the two parts of my argument are first a hard negative argument, addressed to a mechanist putting forward a particular claim, and proving to him, by means he must acknowledge to be valid, that his claim is untenable, and secondly a hand-waving positive argument, addressed to intelligent men, bystanders as well as mechanists espousing particular versions of mechanism, to the effect that some sort of argument on these lines can always be found to deal with any further version of mechanism that may be thought up.

I read the paper to the Oxford Philosophical Society in October 1959 and subsequently published it in *Philosophy*, (1) and later set out the argument in more detail in *The Freedom of the Will*. (2) I have been much attacked. Although I argued with what I hope was becoming modesty and a certain degree of tentativeness, many of the replies have been lacking in either courtesy or caution. I must have touched a raw nerve. That, of course, does not prove that I was right. Indeed, I should at once concede that I am very likely not to be entirely right, and that others will be able to articulate the arguments more clearly, and thus more cogently, than I did. But I am increasingly persuaded that I was not entirely wrong, by reason of the very wide disagreement among my critics about where exactly my arguments fail. Each picks on a different point, allowing that the points objected to by other critics, are in fact all right, but hoping that his one point will prove fatal. None has, so far as I can see. I used to try and answer each point fairly and fully, but the flesh has grown weak. Often I was simply pointing out that the critic was not criticizing any argument I had put forward but one which he would have liked me to put forward even though I had been at pains to discount it. In recent years I have been less zealous to defend myself, and often miss articles altogether. (3) There may be some new decisive objection I have altogether overlooked. But the objections I have come across so far seem far from decisive.

To consider each objection individually would be too lengthy a task to attempt here. I shall pick on five recurrent themes. Some of the objections question the idealisation implicit in the way I set up the contest between the mind and the machine; some raise questions of modality and finitude; some turn on issues of transfinite arithmetic; some are concerned with the extent to which rational inferences should be formalisable; and some are about consistency.

Many philosophers question the idealisation implicit in the Gödelian argument. A context is envisaged between ``the mind'' and ``the machine'', but it is an idealised mind and an idealised machine. Actual minds are embodied in mortal clay; actual machines often malfunction or wear out. Since actual machines are not Turing machines, not having an infinite tape, that is to say an infinite memory, it may be held that they cannot be automatically subject to Gödelian limitations. But Gödel's theorem applies not only to Peano Arithmetic, with its infinitistic postulate of recursive reasoning, but to the weaker Robinson Arithmetic Q, which is only potentially, not actually infinite, and hardly extends beyond the range of plausible computer progress. In any case, limitations of finitude reduce, rather than enhance, the plausibility of some computer's being an adequate representation of a mind. Actual minds are embodied in mortal clay. In the short span of our actual lives we cannot achieve all that much, and might well have neither the time nor the cleverness to work out our Gödelian formula. Hanson points out that there could be a theorem of Elementary Number Theory that I cannot prove because a proof of it would be too long or complex for me to produce. (4) Any machine that represented a mind would be would be enormously complicated, and the calculation of its Gödel sentence might well be beyond the power of any human mathematician. (5) But he could be helped. Other mathematicians might come to his aid, reckoning that they also had an interest in the discomfiture of the mechanical Goliath. (6) The truth of the Gödelian sentence under its intended interpretation in ordinary informal arithmetic is a mathematical truth, which even if pointed out by other mathematicians would not depend on their testimony in the way contingent statements do. So even if aided by the hints of other mathematicians, the mind's asserting the truth of the Gödelian sentence would be a genuine ground for differentiating it from the machine.

Some critics of the Gödelian argument---Dennett, Hofstadter and Kirk---complain that I am insufficiently sensitive to the sophistication of modern computer technology, and that there is a fatal ambiguity between the fundamental level of the machine's operations and the level of input and output that is supposed to represent the mind: in modern parlance, between the machine code and the programming language, such as PROLOG. But although there is a difference of levels, it does not invalidate the argument. A compiler is entirely deterministic. Any sequence of operations specified in machine code can be uniquely specified in the programming language, and vice versa. Hence it is quite fair to characterize the capacity of the mechanist's machine in terms of a higher level language. In order to begin to be a representation of a mind it must be able to do simple arithmetic. And then, at this level, Gödel's theorem applies. The same counter applies to Dennett's complaint that the comparison between men and Turing machines is highly counterintuitive because we are not much given to wandering round uttering obscure truths of ordinary informal arithmetic. Few of us are capable of asserting a Gödelian sentence, fewer still of wanting to do so. ``Men do not sit around uttering theorems in a uniform vocabulary, but say things in earnest and in jest, make slips of the tongue, speak several languages, signal agreement by nodding or otherwise acting nonverbally, and---most troublesome for this account---utter all kinds of nonsense and contradictions, both deliberately and inadvertently." (7) Of course, men are un-machinelike in these ways, and many philosophers have rejected the claims of mechanism on these grounds alone. But mechanists claim that this is too quick. Man, they say, is a very complicated machine, so complicated as to produce all this un-machinelike output. We may regard their contention as highly counter-intuitive, but should not reject it out of hand. I therefore take seriously, though only in order to refute it, the claim that a machine could be constructed to represent the behaviour of a man. If so, it must, among other things, represent a man's mental behaviour. Some men, many men, are capable of recognising a number of basic arithmetical truths, and, particularly when asked to (which can be viewed as a particular input), can assert them as truths. Although ``a characterization of a man as a certain sort of theorem-proving machine" (8) would be a less than complete characterization, it would be an essential part of a characterization of a machine if it was really to represent a man. It would have to be able to include in its output of what could be taken as assertions the basic truths of arithmetic, and to accept as valid inferences those that are validated by first-order logic. This is a minimum. Of course it may be able to do much more---it may have in its memory a store of jokes for use in after-dinner speeches, or personal reminiscences for use on subordinates - but unless its output, for suitable questions or other input, includes a set of assertions itself including Elementary Number Theory, it is a poor representation of some human minds. If it cannot pass O-level maths, are we really going to believe a mechanist when he claims that it represents a graduate?

Actual minds are finite in what they actually achieve. Wang and Boyer see difficulties in the infinite capabilities claimed for the mind as contrasted with the actual finitude of human life. Boyer takes a *post mortem* view, and points out that all of the actual output of Lucas, Astaire, or anyone else can be represented *ex post facto* by a machine. (9) Actual achievements of mortal men are finite, and so simulable. When I am dead it would be possible to program a computer with sufficient graphic capacity to show on a video screen a complete biographical film of my life. But when I am dead it will be easy to outwit me. What is in issue is whether a computer can copy a living me, when I have not as yet done all that I shall do, and can do many different things. It is a question of potentiality rather than actuality that is in issue. Wang concedes this, and allows that we are inclined to say that it is logically possible to have a mind capable of recognising any true proposition of number theory or solving a set of Turing-unsolvable problems, but life is short. (10) In a finite life-span only a finite number of the propositions can be recognised, only a finite set of problems can be solved. And a machine can be programmed to do that. Of course, we reckon that a man *can* go on to do more, but it is difficult to capture that sense of infinite potentiality. This is true. It is difficult to capture the sense of infinite potentiality. But it is an essential part of the our concept of mind, and a modally ``flat" account of the a mind in terms only of what it has done is as unconvincing as an account of cause which considers only constant conjunction, and not what would have been the case had circumstances been different. In order to capture this sense of potentiality, I set out my argument in terms of a challenge which leaves it open to the challenger to meet in any way he likes. Two-sided, or ``dialectical", arguments often succeed in encapsulating concepts that elude explication in purely monologous terms: the epsilon-delta exegesis an infinitesimals is best conveyed thus, and more generally any alternation of quantifiers, as in the EA principles suggested by Professor Clark Glymour for the ultimate convergence of theories on truth.

Although some degree of idealisation seems allowable in considering a mind untrammelled by mortality and a Turing machine with infinite tape, doubts remain as to how far into the infinite it is permissible to stray. Transfinite arithmetic underlies the objections of Good and Hofstadter. The problem arises from the way the contest between the mind and the machine is set up. The object of the contest is not to prove the mind better than the machine, but only different from it, and this is done by the mind's Gödelizing the machine. It is very natural for the mechanist to respond by including the Gödelian sentence in the machine, but of course that makes the machine a different machine with a different Gödelian sentence all of its own, which it cannot produce as true but the mind can. So then the mechanist tries adding a Gödelizing operator, which gives, in effect a whole denumerable infinity of Gödelian sentences. But this, too, can be trumped by the mind, who produces the Gödelian sentence of the new machine incorporating the Gödelizing operator, and out Gödelizes the lot. Essentially this is the move from w (omega), the infinite sequence of Gödelian sentences produced by the Gödelizing operator, to w + 1, the next transfinite ordinal. And so it goes on. Every now and again the mechanist loses patience, and incorporates in his machine a further operator, designed to produce in one fell swoop all the Gödelian sentences the mentalist is trumping him with: this is in effect to produce a new limit ordinal. But such ordinals, although they have no predecessors, have successors just like any other ordinal, and the mind can out-Gödel them by producing the Gödelian sentence of the new version of the machine, and seeing it to be true, which the machine cannot. Hofstadter thinks there is a problem for the mentalist in view of a theorem of Church and Kleene on Formal Definitions of Transfinite Ordinals. (11) They showed that we couldn't program a machine to produce names for all the ordinal numbers. Every now and again some new, creative step is called for, when we consider all the ordinal numbers hitherto named, and we need to encompass them all in a single set, which we can use to define a new sort of ordinal, transcending all previous ones. Hofstadter thinks that, in view of the Church-Kleene theorem, the mind might run out of steam, and fail to think up new ordinals as required, and so fail in the last resort to establish the mind's difference from some machine. But this is wrong on two counts. In the first place it begs the question and in the second it misconstrues the nature of the contest.

Hofstadter assumes that the mind is subject to the same limitations as the machine is, and that since there is no mechanical way of naming all the ordinals, the mind cannot do it either. But this is precisely the point in issue. Gödel himself rejected mechanism on account of our ability to think up fresh definitions for transfinite ordinals (and ever stronger axioms for set theory) and Wang is inclined to do so too.(12) On this occasion, it is pertinent to note that Turing himself was, on this question, of the same mind as Gödel. He was led ``to ordinal logics as a way to `escape' Gödel's incompleteness theorems", (13) but recognised that ``although in pre-Gödel times it was thought by some that it would be able to carry this programmme to such an extent that ... the necessity for intuition would be entirely eliminated," as a result of Gödel's incompleteness theorems one must turn instead to `non-constructive' systems of logic in which ``not all the steps in a proof are mechanical, some being intuitive". Turing concedes that the steps whereby we recognise formulae as ordinal formulae are intuitive, and goes on to say that we should show quite clearly when a step makes use of intuition, and when it is purely formal, and that the strain put on intuition should be a minimum. (14) He clearly, like Gödel, allows that the mind's ability to recognise new ordinals outruns the ability of any formal algorithm to do so, though he does not draw Gödel's conclusion. It may be, indeed, that the mind's ability to recognise new ordinals is the issue on which battle should be joined; Good claimed as much (15) ---though disputes about the notation for ordinals lack the sharp edge of the Gödelian argument. But whatever the merits of different battlefields, it is clear that they are contested areas in the same conflict, and undisputed possession of the one cannot be claimed in order to assert possession of the other.

In any case Hofstadter misconstrues the nature of the contest. All the difficulties are on the side of the mechanist trying to devise a machine that cannot be out-Gödelized. It is the mechanist who resorts to limit ordinals, and who may have problems in devising new notations for them. The mind needs only to go on to the next one, which is always an easy, unproblematic step, and out-Gödelize whatever is the mechanist's latest offering. Hofstadter's argument, as often, tells against the position he is arguing for, and shows up a weakness of machines: there is no reason to suppose that it is shared by minds, and in the nature of the case it is a difficulty for those who are seeking to evade the Gödelian argument, not those who are deploying it.

Underlying Hofstadter's argument is a rhetorical question that many mechanists have raised. ``How does Lucas know that the mind can do this, that, or the other?" It is no good, they hold, that I should opine it or simply assert it; I must prove it. And if I prove it, then since the steps of my proof can be programmed into a machine, the machine can do it too. Good puts the argument explicitly:

What he must prove is that he personally can always make the improvement: it is not sufficient to believe it since belief is a matter of probability and Turing machines are not supposed to be capable of probability judgements. But no such proof is possible since, if it were given, it could be used for the design of a machine that could always do the improving.

The same point is made by Webb in his sustained and searching critique of the Gödelian argument:

It is only because Gödel gives an effective way of constructing the Gödelian sentence that Lucas can feel confident that he can find the Achilles' heel of any machine. But then if Lucas can effectively stump any machine, then there must be a machine which does this too. (16) [This] ``is the basic dilemma confronting anti-mechanism: just when the constructions used in its arguments become effective enough to be sure of, (T) <*viz*. Every humanly effective computation procedure can be simulated by a Turing machine> then implies that a machine can simulate them. In particular it implies that our very behaviour of applying Gödel's argument to arbitrary machines - in order to conclude that we cannot be modelled by a machine - *can indeed be modelled by a machine*. Hence any such conclusion must fail, or else we will have to conclude that certain *machines* cannot be modelled by any machine! In short, anti-mechanist arguments must either be ineffective, or else unable to show that their executor is not a machine." (17)

The core of this argument is an assumption that every informal argument must either be formalisable or else invalid. Such an assumption undercuts the distinction I have drawn between two senses of Gödelian argument: between a negative argument according to an exact specification, which a machine could be programmed to carry out, and on the other hand a certain style of arguing, similar to Gödel's original argument in inspiration, but not completely or precisely specified, and therefore not capable of being programmed into a machine, though capable of being understood and applied by an intelligent mind. Admittedly, we cannot *prove* to a hide-bound mechanist that we can go on. But we may come to a well-grounded confidence that we can, which will give us, and the erstwhile mechanist if he is reasonable and not hide-bound, good reason for rejecting mechanism.

Against this claim of the mentalist that he has got the hang of doing something which cannot be described in terms of a mechanical program, the mechanist says ``Sez you" and will not believe him unless he produces a program showing how he would do it. It is like the argument between the realist and the phenomenalist. The realist claims that there exist entities not observed by anyone: the phenomenalist demands empirical evidence; if it is not forthcoming, he remains sceptical of the realist's claim; if it is, then the entity is not unobserved. In like manner the mechanist is sceptical of the mentalist's claim unless he produces a specification of how he would do what a machine cannot: if such a specification is not forthcoming, he remains sceptical; if it is, it serves as a basis for programming a machine to do it after all. The mechanist position, like the phenomenalist, is invulnerable but unconvincing. I cannot prove to the mechanist that anything can be done other than what a machine can do, because he has restricted what he will accept as a proof to such an extent that only ``machine-doable" deeds will be accounted doable at all. But not all mechanists are so limited. Many mechanists and many mentalists are rational agents wondering whether in the light of modern science and cybernetics mechanism is, or is not, true. They have not closed their minds by so redefining proof that none but mechanist conclusions can be established. They can recognise in themselves their having ``got the hang" of something, even though no program can be written for giving a machine the hang of it. The parallel with the *Sorites* argument is helpful. Arguing against a finitist, who does not accept the principle of mathematical induction, I may see at the meta-level that if he has conceded F(0) and (Ax)(F(x) --> F(x + 1)) then I can claim without fear of contradiction (Ax)F(x). I can be quite confident of this, although I have no finitist proof of it. All I can do, *vis à vis* the finitist, is to point out that *if* he were to deny my claim in any specific instance, I could refute him. True, a finitist could refute him too. But I have generalised in a way a finitist could not, so that although each particular refuting argument is finite, the claim is infinite. In a similar fashion each Gödelian argument is effective, and will convince even the mechanist that he is wrong; but the generalisation from individual tactical refutations to a strategic claim does not have to be effective in the same sense, although it may be entirely rational for the mind to make the claim.

Nevertheless an air of paradox remains. The idea of a totally intuitive, unformalisable argument arouses suspicion: if it can convince, it can be conveyed, and if it can be conveyed, it can be formulated and expressed in formal terms. Let me therefore stress that I am not claiming that my, or any, argument is absolutely unformalisable. Any argument can be formalised, as the Tortoise proved to Achilles, the formal axiom or rule of inference invoked will be no more convincing than the original unformalised argument. I am not claiming that the Gödelian argument cannot be formalised, but that whatever formalisation we adopt, there are further arguments which are clearly valid but not captured by that formalisation. Not only, again as the Tortoise proved to Achilles, must we always be ready to recognise some rules of inference as applying and inferences as valid without more ado, but we shall be led, if we are rational, to extend our range of acknowledged valid inferences beyond any antecedently laid down bounds. This does not preclude our subsequently formalising them, only our supposing that any formalisation is inferentially complete.

But we always can formalise; in particular, we can formalise the argument that Gödel uses to prove that the Gödelian formula is unprovable-in-the-system but none the less true. At first sight there seems to be a paradox. Gödel's argument purports to show that the Gödelian sentence is unprovable but true. But if it shows that the Gödelian sentence is true, surely it has proved it, so that it is provable after all. The paradox in this case is resolved by distinguishing provability-in-the-formal-system from the informal provability given by Gödel's reasoning. But this reasoning can be formalised. We can go over Gödel's argument step by step, and formalise it. If we do so we find that an essential assumption for his argument that the Gödelian sentence is unprovable is that the formal system should be consistent. Else every sentence would be provable, and the Gödelian sentence, instead of being unprovable and therefore true, could be provable and false. So what we obtain, if we formalise Gödel's informal argumentation, is not a formal proof within Elementary Number Theory (ENT for short) that the Gödelian sentence, G is true, but a formal proof within Elementary Number Theory

|- Cons(ENT) --> G

where Cons(ENT) is a sentence expressing the consistency of Elementary Number Theory. Only if we also had a proof in Elementary Number Theory yielding

|- Cons(ENT)

would we be able to infer by *Modus Ponens*

|- G

Since we know that

¬ |- G, [i.e. G is not derivable: this is the best I can do to render symbolic logic in HTML]

we infer also that

¬:|- Cons(ENT). [i.e. Cons(ENT) is not derivable]

This is Gödel's second theorem. Many critics have appealed to it in order to fault the Gödelian argument. Only if the machine's formal system is consistent and we are in a position to assert its consistency are we really able to maintain that the Gödelian sentence is true. But we have no warrant for this. For all we know, the machine we are dealing with may be inconsistent, and even if it is consistent we are not entitled to claim that it is. And in default of such entitlement, all we have succeeded in proving is

|- Cons(ENT) --> G,

and the machine can do that too.

These criticisms rest upon two substantial points: the consistency of the machine's system *is* assumed by the Gödelian argument and *cannot* be always established by a standard decision-procedure. The question ``By what right does the mind assume that the machine is consistent?" is therefore pertinent. But the moves made by mechanists to deny the mind that knowledge are unconvincing. Paul Benacerraf suggests that the mechanist can escape the Gödelian argument by not staking out his claim in detail. (18) The mechanist offers a ``Black Box" without specifying its program, and refusing to give away further details beyond the claim that the black box represents a mind. But such a position is both vacuous and untenable: vacuous because there is no content to mechanism unless some specification is given---if I am presented with a black box but ``told not to peek inside" then why should I think it contains a machine and not, say, a little black man? The mechanist's position is also untenable: for although the mechanist has refused to specify what machine it is that he claims to represent the mind, it is evident that the Gödelian argument would work for any consistent machine and that an inconsistent machine would be an implausible representation. The stratagem of playing with his cards very close to his chest in order to deny the mind the premisses it needs is a confession of defeat.

Putnam contends that there is an illegitimate inference from the true premiss

I can see that (Cons(ENT) ---> G)

to the false conclusion

Cons(ENT) --> I can see that (G). (19)

It is the latter that is needed to differentiate the mind from the machine, for what Gödel's theorem shows is

Cons(ENT) ---> ENT machine cannot see that (G),

but it is only the former, according to Putnam, that I am entitled to assert. Putnam's objection fails on account of the dialectical nature of the Gödelian argument. The mind does not go round uttering theorems in the hope of tripping up any machines that may be around. Rather, there is a claim being seriously maintained by the mechanist that the mind can be represented by some machine. Before wasting time on the mechanist's claim, it is reasonable to ask him some questions about his machine to see whether his seriously maintained claim has serious backing. It is reasonable to ask him not only what the specification of the machine is, but whether it is consistent. Unless it is consistent, the claim will not get off the ground. If it is warranted to be consistent, then that gives the mind the premiss it needs. The consistency of the machine is established not by the mathematical ability of the mind but on the word of the mechanist. The mechanist has claimed that his machine is consistent. If so, it cannot prove its Gödelian sentence, which the mind can none the less see to be true: if not, it is out of court anyhow.

Wang concedes that it is reasonable to contend that only consistent machines are serious candidates for representing the mind, but then objects it is too stringent a requirement for the mechanist to meet because there is no decision-procedure that will always tell us whether a formal system strong enough to include Elementary Number Theory is consistent or not. (20) But the fact that there is no decision-*procedure* means only that we cannot always tell, not that we can never tell. Often we can tell that a formal system is not consistent---*e.g.* it proves as a theorem:

|- p&¬p

or,

|- 0 = 1

Also, *we* may be able to tell that a system *is* consistent. We have finitary consistency proofs for propositional calculus and first-order predicate calculus, and Gentzen's proof, involving transfinite induction, for Elementary Number Theory. We are therefore not asking the impossible of the mechanist in requiring him to do some preliminary sorting out before presenting candidates for being plausible representations of the mind. Unless they satisfy the examiner---the mechanist---in Prelims on the score of consistency, they are not eligible to enter for Finals, and all those that are thus qualified can be sure of failing for not being able to assert their Gödelian sentence.

The two-stage examination is thus able to sort out the inconsistent sheep who fail the qualifying examination from the consistent goats who fail their finals, and hence enables us to take on all challenges even from inconsistent machines, without pretending to possess superhuman powers. Although all machines are entitled to enter for the mind-representation examination, only relatively few machines are plausible candidates for representing the mind, and there is no need to take a candidate seriously just because it is a machine. If the mechanist's claim is to be taken seriously, some recommendation will be required, and at the very least a warranty of consistency would be essential. Wang protests that this is to expect superhuman powers of him, and in a response to Benacerraf's ``God, The Devil and Gödel'', I picked up his suggestion that the mechanist might be no mere man but the Prince of Darkness himself to whom the question of whether the machine was consistent or not could be addressed in expectation of an answer. (21) Rather than ask high-flown questions about the mind we can ask the mechanist the single question whether or not the machine that is proposed as a representation of the mind would affirm the Gödelian sentence of its system. If the mechanist says that his machine will affirm the Gödelian sentence, the mind then will know that it is inconsistent and will affirm anything, quite unlike the mind which is characteristically selective in its intellectual output. If the mechanist says that his machine will not affirm the Gödelian sentence, the mind then will know since there was at least one sentence it could not prove in its system it must be consistent; and knowing that, the mind will know that the machine's Gödelian sentence is true, and thus will differ from the machine in its intellectual output. And if the mechanist is merely human, and moreover does not know what answer the machine would give to the Gödelian question, he has not done his home-work properly, and should go away and try to find out before expecting us to take him seriously.

In asking the mechanist rather than the machine, we are making use of the fact that the issue is one of principle, not of practice. The mechanist is not putting forward actual machines which actually represent some human being's intellectual output, but is claiming instead that there could in principle be such a machine. He is inviting us to make an intellectual leap, extrapolating from various scientific theories and skating over many difficulties. He is quite entitled to do this. But having done this he is not entitled to be coy about his in-principle machine's intellectual capabilities or to refuse to answer embarrassing questions. The thought-experiment, once undertaken, must be thought through. And when it is thought through it is impaled on the horns of a dilemma. Either the machine can prove in its system the Gödelian sentence or it cannot: if it can, it is inconsistent, and not equivalent to a mind; if it cannot, it is consistent, and the mind can therefore assert the Gödelian sentence to be true. Either way the machine is not equivalent to the mind, and the mechanist thesis fails.

A number of thinkers have chosen to impale themselves on the inconsistency horn of the dilemma. We are machines, they say, but very limited, fallible and inconsistent ones. In view of our many contradictions, changes of mind and failures of logic, we have no warrant for supposing the mind to be consistent, and therefore no ground for disqualifying a machine for inconsistency as a candidate for being a representation of the mind. Hofstadter thinks it would be perfectly possible to have an artificial intelligence in which propositional reasoning emerged as consequences rather than as being pre-programmed. ``And there is no particular reason to assume that the strict Propositional Calculus, with its rigid rules and the rather silly definition of consistency they entail, would emerge from such a program." (22)

None of these arguments goes any way to making an inconsistent machine a plausible representation of a mind. Admittedly the word `consistent' is used in different senses, and the claim that a mind is consistent is likely to involve a different sense of consistency and to be established by different sorts of arguments from those in issue when a machine is said to be consistent. If this is enough to establish the difference between minds and machines, well and good. But many mechanists will not be so quickly persuaded and will maintain that a machine can be programmed, in some such way as Hofstadter supposes, to emit mind-like behaviour. In that case it is machine-like consistency rather than mind-like consistency that is in issue. Any machine, if it is to begin to represent the output of a mind must be able to operate with symbols that can be plausibly interpreted as negation, conjunction, implication, *etc*., and so must be subject to the rules of some variant of the propositional calculus. Unless something rather like the propositional calculus with some comparable requirement of consistency emerges from the program of a machine, it will not be a plausible representation of a mind, no matter no matter how good it is as a specimen of Artificial Intelligence. Of course, any plausible representation of a mind would have to manifest the behaviour instanced by Wang, constantly checking whether a contradiction had been reached and attempting to revise its basic axioms when that happened. But this would have to be in accordance with certain rules. There would have to be a program giving precise instructions how the checking was to be undertaken, and in what order axioms were to be revised. Some axioms would need to be fairly immune to revision. Although some thinkers are prepared to envisage a logistic calculus in which the basic inferences of propositional calculus do not hold (*e.g.* from p & q to p) or the axioms of Elementary Number Theory have been rejected, any machine which resorted to such a stratagem to avoid contradiction would also lose all credence as a representation of a mind. Although we sometimes contradict ourselves and change our minds, some parts of our conceptual structure are very stable, and immune to revision. Of course it is not an absolute immunity. One can allow the Cartesian possibility of conceptual revision without being guilty, as Hutton supposes, (23) of inconsistency in claiming knowledge of his own consistency. To claim to know something is not to claim infallibility but only to have adequate backing for what is asserted. Else all knowledge of contingent truths would be impossible. Although one cannot say `I know it, although I *may* be wrong', it is perfectly permissible to say `I know it, although I *might conceivably* be wrong'. So long as a man has good reasons, he can responsibly issue a warranty in the form of a statement that he knows, even though we can conceive of circumstances in which his claim would prove false and would have to be withdrawn. So it is with our claim to know the basic parts of our conceptual structure, such as the principles of reasoning embodied in the propositional calculus or the truths of ordinary informal arithmetic. We have adequate, more than adequate, reason for affirming our own consistency and the truth, and hence also the consistency, of informal arithmetic, and so can properly say that we know, and that any machine representation of the mind must manifest an output expressed by a formal (since it is a machine) system which is consistent and includes Elementary Number Theory (since it is supposed to represent the mind). But there remains the Cartesian possibility of our being wrong, and that we need now to discuss. Some mechanists have conceded that a consistent machine could be out-Gödeled by a mind, but have maintained that the machine representation of the mind is an inconsistent machine, but one whose inconsistency is so deep that it would take a long time ever to come to light. It therefore would avoid the quick death of non-selectivity. Although in principle it could be brought to affirm anything, in practice it will be selective, affirming some things and denying others. Only in the long run will it age---or mellow, as we kindly term it---and then ``crash" and cease to deny anything; and in the long run we die---usually before suffering senile dementia. Such a suggestion chimes in with a line of reasoning which has been noticeable in Western Thought since the Eighteenth Century.

Reason, it is held, suffers from certain antinomies, and by its own dialectic gives rise to internal contradictions which it is quite powerless to reconcile, and which must in the end bring the whole edifice crashing down in ruins. If the mind is really an inconsistent machine then the philosophers in the Hegelian tradition who have spoken of the self-destructiveness of reason are simply those in whom the inconsistency has surfaced relatively rapidly. They are the ones who have understood the inherent inconsistency of reason, and who, negating negation, have abandoned hope of rational discourse, and having brought mind to the end of its tether, have had on offer only counsels of despair.

Against this position the Gödelian argument can avail us nothing. Quite other arguments and other attitudes are required as antidotes to nihilism. It has long been sensed that materialism leads to nihilism, and the Gödelian argument can be seen as making this *reductio* explicit. And it is a *reductio*. For mechanism claims to be a rational position. It rests its case on the advances of science, the underlying assumptions of scientific thinking and the actual achievements of scientific research. Although other people may be led to nihilism by feelings of *angst* or other intimations of nothingness, the mechanist must advance arguments or abandon his advocacy altogether. On the face of it we are not machines. Arguments may be adduced to show that appearances are deceptive, and that really we are machines, but arguments presuppose rationality, and if, thanks to the Gödelian argument, the only tenable form of mechanism is that we are inconsistent machines, with all minds being ultimately inconsistent, then mechanism itself is committed to the irrationality of argument, and no rational case for it can be sustained.

Notes

(*) A paper read to the Turing Conference at Brighton on April 6th, 1990 by J.R. Lucas Fellow of Merton College, Oxford. For the copyright of the papers of J.R. Lucas see http://users.ox.ac.uk/~jrlucas/ back

(1) Minds, Machines and Gödel, *Philosophy*, **36**, 1961, pp.112-127; reprinted in Kenneth M.Sayre and Frederick J.Crosson, eds., *The Modeling of Mind*, Notre Dame, 1963, pp. 255-271; and in A.R.Anderson, *Minds and Machines*, Prentice-Hall, 1964, pp. 43-59. back

(2) *The Freedom of the Will*, Oxford, 1970 (now avaialble again). back

(3) I give at the end a list of some of the major criticisms I have come across. back

(4) William Hanson, ``Mechanism and Gödel's Theorems," *British Journal for the Philosophy of Science*, XXII, 1971, p.12; compare Hofstadter, 1979, p.475. back

(5) Rudy Rucker, ``Gödel's Theorem: The Paradox at the heart of modern man", *Popular Computing*, February 1985, p.168. back

(6) I owe this suggestion to M.A.E. Dummett, at the original meeting of the Oxford Philosophical Society on October 30th, 1959. A similar suggestion is implicit in Hao Wang, *From Mathematics to Philosophy*, London, 1974, p.316. back

(7) D.C.Dennett, Review of *The Freedom of the Will*, in *Journal of Philosophy*, 1972, p.530. back

(8) P.527. back

(9) David L.Boyer, ``Lucas, Gödel and Astaire", *The Philosophical Quarterly*, 1983, pp. 147-159. back

(10) Hao Wang, *From Mathematics to Philosophy*, London, 1974, p.315. back

(11) Douglas R.Hofstadter, *Gödel, Escher, Bach*, New York, 1979, p.475. back

(12) Hao Wang, *From Mathematics to Philosophy*, London, 1974, pp.324-326. back

(13) Solomon Feferman, "Turing in the Land of O(z)", in Rolf Herken ed., *The Universal Turing Machine*, Oxford, 1988, p.121. back

(14) A.M.Turing, "Systems of logic based on ordinals", Proceedings of the London Mathematical Society, (2), 45, 1939, pp.161-228; reprinted in M.Davis, *The Undecidable*, New York, 1965; quoted by Solomon Feferman, op.cit., p.129. back
(15) I.J.Good, ``Gödel's Theorem is a Red Herring", *British Journal for the Philosophy of Science*, **19**, 1968, pp. 357-358. back
(16) Judson C.Webb, *Mechanism, Mentalism and Metamathematics; An Essay on Finitism*, Dordrecht, 1980, p.230. back
(17) P.232, Webb's italics. back
(18) Paul Benacerraf, God, The Devil and Gödel, *The Monist*, **51**, 1967, pp. back
(19) Hilary Putnam ``Minds and Machines", in Sidney Hook, ed., *Dimensions of Mind: A Symposium*, New York, 1960; reprinted in Kenneth M. Sayre and Frederick J.Crosson, eds., *The Modeling of Mind*, Notre Dame, 1963, pp. 255-271; and in A. R. Anderson, *Minds and Machines*, Prentice-Hall, 1964, pp. 43-59. back
(20) Hao Wang, *From Mathematics to Philosophy*, London, 1974, p.317. back
(21) Paul Benacerraf, God, The Devil and Gödel, *The Monist*, **51**, 1967, pp. 22-23; J.R. Lucas, "Satan Stultified", The Monist, 52, 1967, pp. 152-3. back
(22) Hofstadter, 1979, p.578; cf. Charles S. Chihara, ``On Alleged Refutations of Mechanism using Gödel's Incompleteness Results", *Journal of Philosophy*, LXIX, no.17, 1972, p.526. back
(23) Anthony Hutton, This Gödel is Killing Me, *Philosophia*, vol. 6, no.1, 1976, pp. 135-144. back

Criticisms of the Gödelian Argument

J.J.C.Smart, ``Gödel's Theorem, Church's Theorem, and Mechanism", *Synthese*, **13**, 1961.
J.J.C.Smart, ``Man as a Physical Mechanism", ch.VI of his *Philosophy and Scientific Realism*.
Hilary Putnam ``Minds and Machines", in Sidney Hook, ed., *Dimensions of Mind. A Symposium*, New York, 1960; reprinted in Kenneth M. Sayre and Frederick J. Crosson, eds., *The Modeling of Mind*, Notre Dame, 1963, pp. 255-271; and in A. R. Anderson, *Minds and Machines*, Prentice-Hall, 1964, pp. 43-59.
C.H. Whitely, ``Minds, Machines and Gödel: a Reply to Mr. Lucas", *Philosophy*, **37**, 1962, pp.61-62.
Paul Benacerraf, God, the Devil and Gödel, *The Monist*, 1967, pp. 9-32.
I.J. Good, Human and Machine Logic, *British Journal for the Philosophy of Science*, **18**, 1967, pp. 144-147.
I.J.Good, ``Gödel's Theorem is a Red Herring", *British Journal for the Philosophy of Science*, **19**, 1968, pp. 357-8.
David Lewis, Lucas Against Mechanism, *Philosophy*, XLIV, 1969, pp. 231-233.
David Coder, ``Goedel's Theorem and Mechanism", *Philosophy*, XLIV, 1969, pp. 234-237, esp. p.236.
Jonathan Glover, *Responsibility*, London, 1970, p.31.
William Hanson, ``Mechanism and Gödel's Theorems," *British Journal for the Philosophy of Science*, XXII, 1971.
D.C. Dennett, Review of *The Freedom of the Will*, *Journal of Philosophy*, 1972.
Charles S. Chihara, ``On Alleged Refutations of Mechanism using Gödel's Incompleteness Results", *Journal of Philosophy*, LXIX, no.17, 1972.
Hao Wang, *From Mathematics to Philosophy*, London, 1974, pp.319, 320, 324-326.
A.J.P.Kenny in A.J.P.Kenny, H.C.Longuet-Higgins, J.R. Lucas and C.H.Waddington, *The Nature of Mind*, Edinburgh, 1976, p.75.
Anthony Hutton, ``This Gödel is Killing Me", *Philosophia*, vol. 6, no.1, 1976, pp. 135-144.
J.W. Thorp, ``Free Will and Neurophysiological Determinism", Oxford D.Phil. Thesis, 1976, p.79.

J.L. Mackie, *Ethics: Inventing Right and Wrong*, Penguin, 1977, p. 219.
David Lewis, ``Lucas Against Mechanism II", *Canadian Journal of Philosophy,* IX, 1979, pp. 373-376.
Douglas R.Hofstadter, *Gödel, Escher, Bach,* New York, 1979, p.475.
Emmanuel Q. Fernando, ``Mathematical and Philosophical Implications of the Gödel Incompleteness Theorems". M.A. Thesis, College of Arts and Sciences, University of the Philippines, Quezu City, September 1980.
Judson C. Webb, *Mechanism, Mentalism and Metamathematics; An Essay on Finitism*, Dordrecht, 1980, p.230.
G. Lee Bowie, ``Lucas' Number is Finally Up", *Journal of Philosophical Logic*, **11**, 1982, pp.279-285.
P.Sleazak, ``Gödel's Theorem and the Mind", *British Journal for the Philosophy of Science*, XXXIII, 1982.
Rudy Rucker, ``Gödel's Theorem: The Paradox at the heart of modern man", *Popular Computing*, February 1985, p.168.
David L. Boyer, ``Lucas, Gödel and Astaire", *The Philosophical Quarterly*, 1983, pp.147-159.
David Bostock, ``Gödel and Determinism", private communication, November, 1984.
Robert Kirk, ``Mental Machinery and Gödel", *Synthese*, **66**, 1986, pp.437-452.
Other works are cited in *The Freedom of the Will*, pp. 174-6.