Etica & Politica / Ethics & Politics, 2003, 1
http://www.units.it/~dipfilo/etica_e_politica/2003_1/13_monographica.htm

The Gödelian Argument: Turn Over the Page (*)

J.R. Lucas

Fellow of Merton College, Oxford
Fellow of the British Academy

I want to start by quarrelling with Sir Roger Penrose. In 1990 the *Journal of Behavioral and Brain Sciences* published a large number of peer reviews of his book, The Emperor's New Mind. At the end he said in his response:
All my adverse critics on this topic have jumped to conclusions and, in one way or another, have missed the point of what I am trying to say. None seem to have grasped the full import of the Gödelian argument. The fault is mine: I should have explained things more clearly. (1)
I have no quarrel with the first two sentences: but the third, though charitable and courteous, is quite untrue. Although there are criticisms which can be levelled against the Gödelian argument, most of the critics have not read either of my, or either of Penrose's, expositions carefully, and seek to refute arguments we never put forward, or else propose as a fatal objection one that had already been considered and countered in our expositions of the argument. Hence my title. The Gödelian Argument uses Gödel's theorem to show that minds cannot be explained in purely mechanist terms. It has been put forward, in different forms, by Gödel himself, by Penrose, and by me.
Gödel gave the Gibbs lecture on Boxing Day, 1951, twenty years after he had discovered his theorem, to an audience in the United States, but the lecture was not published or much known about, until after his death. It appeared in the third volume of his Collected Works, which was published in 1995. I read a paper, "Minds, Machines and Gödel" to the Oxford Philosophical Society on October, 30th, 1959, which was subsequently published in Philosophy, 36, 1961, pp.112- 127, and 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. In 1970 I published a fuller version in my The Freedom of the Will, in which I went in greater detail into objections to the Gödelian argument and how they should be answered. Roger Penrose had been thinking about the problem for many years before he published his The Emperor's New Clothes in 1989, which attracted much attention. In 1994 he published Shadows of the Mind in which he countered some of the criticisms that had been levelled against the earlier version of the argument. There have been a large number of discussions, mostly critical, of both his and my versions of the argument.
Gödel argues for a disjunction: an Either/Or, with the strong suggestion that the second disjunct is untenable, and hence by Modus Tollendo Ponens that the first disjunct must be true.
So the following disjunctive conclusion is inevitable: Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified . . . (2)
It is clear that Gödel thought the second disjunct false, so that by Modus Tollendo Ponens he was implicitly denying that any Turing machine could emulate the powers of the human mind. (3)

Roger Penrose uses not Gödel's theorem itself but one of its corollaries, Turing's theorem, which he applies to the whole world-wide activity of all mathematicians put together, and claims that their creative activity cannot be completely accounted for by any algorithm, any set of rigid rules that a Turing machine could be programmed to follow.

I used Gödel's theorem itself, and considered only individuals, reasonably numerate (able to follow and understand Gödel's theorem) but not professional mathematicians. I did not give a direct argument, but rather a schema, a schema of disproof, whereby any claim that a particular individual could be adequately represented by a Turing machine could be refuted. My version was, designedly, much less formal than the others, partly because I was addressing a not-very-numerate audience, but chiefly because I was not giving a direct disproof, but rather a schema which needed to be adapted to refute the particular case being propounded by the other side. I was trying to convey the spirit of disproof, not to dot every i and cross every t, which might have worked against one claim but would have failed against others. I also chose, in arguing against the thesis that the mind can be represented by a Turing machine, to use Gödel's theorem, not Turing's. This was in part because, once again, Gödel's theorem is easier to get the flavour of than Turing's theorem, but also because it involves the concept of truth, itself a peculiarly mental concept.

Other differences need to be noted. Gödel was a convinced dualist. He thought it obvious that minds were essentially different from, and irreducible to, matter; one reason, perhaps, why he did not make more of his argument was that he did not feel the need to refute materialism: why waste effort flogging a dead horse? Penrose is a materialist, but thinks that physics needs to be radically revised in order to accommodate mental phenomena. Quite apart from this, he reckons physics must be developed to account for the phenomenon of the collapse of the psi function in quantum mechanics. He hopes to produce a unified theory which will be both a non-algorithmic theory of quantum collapse and accommodate the phenomenon of mind. I acknowledge the importance of both these problems, but think they are separate. Instead of trying to expand physics in order to have a physical theory of mind, I distinguish sharply two different types of explanation, the regularity explanations we use to explain natural phenomena and the rational explanations we use to justify and explain the actions of rational agents. In that sense I am a dualist. I have difficulty with the full-blooded Cartesian dualism of different sorts of substance, which was, I think, Gödel's position, and am, as regards substance only a one-and-a-halfist at most. But, as regards explanation, I am at least a dualist, indeed, a many- times-more-than-dualist.

Many objections have been raised against the Gödelian argument. Many AI enthusiasts protest that they do not work with Turing machines, and have much more complicated and subtle connexionist systems. I do not dispute that, and specifically allowed in my original article that we might one day be able to create something of silicon with a mind of its own, just as we are able now to procreate carbon based bodies with minds of their own. (4) To those who say I am therefore flogging a dead horse, I reply by citing Breuel, who explains

the reason AI researchers, for practical purposes, adhere to the idea that brains are no more than computational devices is not philosophical stubbornness but the fact that no physical process is known to exist that can be used to build a device computationally more powerful than a Turing machine, and no concrete theories of psychological and cognitive phenomena have so far required any recourse to physical mechanisms that were more powerful than a Turing machine. (5)

We are dealing not only with practical attempts to build machines that can insert a collar stud or tie a bow tie, but with attempts to understand the workings of the human mind; and then to show that one very widely accepted schema of explanation is unavailable is well worth doing.

A much more serious objection is based on the assumption of consistency that I make. Hilary Putnam, when I first put the argument to him in a bar in Princeton, objected that in order to be in a position to know that the Gödelian formula was true, one needed to know that the system

was consistent; he maintained that Gödel's second theorem showed that this was impossible. (6) Many other critics have maintained the same:

Thus the premise that the Gödel sentence is true (and unprovable) cannot be known unless it is known that arithmetic is consistent, and no Turing machine can know the latter. So who says that humans can know it either: (7)

Well, Gentzen for a start. He gave a convincing proof of the consistency of Peano Arithmetic (the simple arithmetic of the natural numbers), using transfinite induction. What Gödel's Second theorem showed was that a system could not be proved to be consistent within itself: we cannot prove Peano Arithmetic consistent from the axioms and by means of just the rules of Peano Arithmetic, but we can give a consistency proof---a very convincing one---applying principles from outside Peano Arithmetic. Such proofs, of course, depend on the wider set of principles being consistent, and that assumption can be called into question, the more particularly since Russell's pointing out that Frege's set theory was inconsistent. (8)

These are proper objections, but not insuperable ones. Although by Gödel's second theorem we cannot prove the consistency of a formal system within that formal system, we can argue for the consistency of a formal system by means of wider considerations. And so when Putnam raised his objection, I countered that although a Turing machine could not, without inconsistency, prove its own consistency, we could affirm that any plausible representation of a mind must be consistent, since minds were selective and were unwilling to assert anything whatsoever, which an inconsistent machine will do. I discussed the matter both then and in my article and book, (9) but Putnam in his review of Shadows of the Mind in the New York Times Book Review (***, p.7) ignores the argument, and says simply:

Mr Lucas's mistake was to confuse two very different statements that could be called "the statement that S is consistent." In particular Mr Lucas confused the colloquial statement that the methods of mathematicians use cannot lead to inconsistent results with the very complex mathematical statement that would arise if we were to apply Gödel's theorem to a hypothetical formalization of these methods.

But I did not confuse them. I argued in some detail, considering and countering various objections that might be put forward, that the two were connected (why else would the word "consistent" be applied in each case?), and that if a Turing machine was inconsistent it could not be a plausible representation of the mind.

Since many critics are unaware of the argument, (10) and are unlikely to look back at papers published some time ago, it is worth articulating the argument afresh. Here in Oxford in the fifth week of Trinity Term it is useful to borrow the terminology of First and Second Public Examinations. The Mechanist claims to have a model of the mind. We ask him whether it is consistent: if he cannot vouch for its consistency, it fails at the first examination; it just does not qualify as a plausible representation, since it does not distinguish those propositions it should affirm from those that it should deny, but is prepared to affirm both undiscriminatingly. We take the Mechanist seriously only if he will warrant that his purported model of the mind is consistent. In that case it passes the First Public Examination, but comes down at the Second, because knowing that it is consistent, we know that its Gödelian formula is true, which it cannot itself produce as true. More succinctly, we can, if a Mechanist presents us with a system that he claims is a model of the mind, ask him simply whether or not it can prove its Gödelian formula (according to some system of Gödel numbering). If he says it can, we know that it is inconsistent, and would be equally able to prove that 2 and 2 make 5, or that 0=1, and we waste little time on examining it. If, however, he acknowledges that the system cannot prove its Gödelian formula, then we know it is consistent, since it cannot prove every well-formed formula, and knowing that it is consistent, know also that its Gödelian formula is true.

In this formulation we have, essentially, a dialogue between the Mechanist and the Mentalist, as we may call him, with the Mechanist claiming to be able to produce a mechanist model of the Mentalist's mind, and the Mentalist being able to refute each particular instance offered.

Many critics have failed to note the dialectical character of the argument, and have rushed in to show that the Mentalist is not in all respects superior to all minds, but has his own limitations, and can often be beaten by a mind. But if only they had turned over the page, they would have seen that I acknowledged as much, and was not making a general claim of superiority, but only a particular one of some difference in each particular case. Other critics try to avoid the dangerous dialogue by having the mechanist not show his hand. (11) He does not tell us which precise model of Turing machine represents a particular mind, nor whether or not the purported mechanist model is consistent; and raises the question whether I, or any human mind, could really fathom the immense complexity of a representation of a human brain. But then why should I? It is for the Mechanist to make good his case. I cannot be just an abstract idea of a turing machine, a generalised Turing machine, I know not what. I must be a particular definite machine. Although Benacerraf may plead ignorance, it must in principle be knowable which machine it is that purports to represent me, and whether it is consistent. And then the argument proceeds.

But still, it may be objected, the Turing machine will be fiendishly complicated. Presented with an enormous printout of gobbledygok, how could I make out what it meant, or what its Gödelian formula was? But it does not have to me just me. As Michael Dummett pointed out when I read the original paper to the Oxford Philosophical Society, I could be helped, indeed helped by all the mathematicians in the world, who might be keen to see a mind, even mine, defeat mechanism. (In this point there is a similarity with Penrose's argument, which is concerned with the output of the entire community of working mathematicians.) Also, now, of course, I could be helped by computers. Not everything has to be in machine code: we can have programs to translate into higher-level, more transparent languages. It would, admittedly, still be difficult to identify all the transformation rules, all the inferences that the machine could make, all its initial assumptions, but not impossibly so, if the mind really were a machine, and really did proceed according to some algorithmic rules. And once we had done this, and chosen some suitable scheme of Gödel numbering, we could set about calculating what the Gödelian formula for that system under that scheme of Gödel numbering must be.

But, it is sometimes further objected, in order to be confident that I can always calculate the Gödelian formula, I must have an algorithm to do it by, in which case a machine could be programmed to do it too. (12) But this is not so. It is easiest seen if we pursue a criticism of I.J. Good, who pressed home an objection I had countered, that the Mechanist might incorporate a Gödelizing operator, which indeed he could, but at the cost of making the machine a different machine and hence with a different Gödel formula. But the move could be iterated, and though at each stage the machine would be different, we should be engaged in a game of catch-as-catch-can in the transfinite ordinals. (13) And there is no algorithm for naming the transfinite ordinals. (14) Every now and again we run out of existing names, and have to devise something new. I claim that this is something we can be confident of doing. My critics claim that I am being over-confident, and that only if there were an algorithm would I be justified in maintaining that I always could go on, and there is no such algorithm. But, say I, we do not have to have an algorithm; when we run out of standard names of transfinite ordinals, we invent new ones, not according to some rule but by the exercise of ingenuity. And, more generally, once we have got the hang of the Gödelian argument, we can adapt it as necessary to the needs of the particular case, and can go on producing appropriate Gödelian formulae, improvising as necessary when the going is not entirely straightforward.

At this stage the importance of originality and creativity begins to emerge. What I have offered is not one knock-down argument but a schema of refutation, which can be seen to work in some cases, but needs adaptation to the particular case. When we consider not a particular case, but all possible cases, I cannot offer a single, all-sweeping argument, but only an approach, relying on the mind's ability to improvise as needed in new circumstance. It is a difficult schema of argumentation, and critics often try to reconstrue it as if it were a single all-

encompassing argument, and then find their reconstruction faulty. In spite of my specific disclaimer, (15) they suppose that I am trying to prove that the mind is better than all computers taken together. And then it is easy to point out that anything I do can be simulated by a computer suitably programmed. (16) All I can do is to repeat my original argument, that I do not claim to be better than all computers, but can show for each particular one that I am different from it. It is the Mechanist's claim that is being evaluated, and can be outwitted in each particular case and shown by the mind it claims to represent, to be an inadequate representation of that very mind. (17)

The argument bifurcates. The Mechanists refuse to acknowledge any originality of the mind. And I cannot make them. If the only thing that will budge the Mechanists is a rule-governed inference which cannot be resisted on pain of inconsistency, then they cannot be made to see the general applicability of Gödelian arguments. All that can be done is to refute each and every particular claim they put forward. But their obduracy is unappealing. Once we get the hang of the Gödelian argument, we see that it will be applicable in all cases, though its mode of application will have to be altered to fit each case individually. There is a way of arguing that commends itself to those possessed of minds, who get the hang of the Gödelian argument, and twig that they can apply it, suitably adapted, in each and every case that crops up. Mechanists may refuse to see the general case, and, acknowledging only knock-down arguments, will have to be knocked down each time they put forward a detailed case: minds can generalise, and will realise that defeat for the Mechanists is always inevitable.

The originality required of the mind is not very great, and it may still be objected that the Gödelian argument has not done much to vindicate real creativity in the face of sceptical doubt. So far as the argument thus far adduced, it is a fair complaint. All that has been achieved so far is to ``defeat the defeaters''; a line of argument, supposedly supported by the success of science, which would lead to a reductionist view of the mind, and the elimination of all originality, has been defeated. We no longer have to think of the mind as an essentially dull automaton, but may still observe that many minds are nonetheless somewhat dull, not to say boring. The Gödelian argument may refute mechanism, but leaves the woodenness of ordinary, uninspired human life untouched. (18) But, although Gödel cannot make us scintillate, he does show that scintillation is conceptually possible. He shows us that to be reasonable is not necessarily to be rule-governed, and that actions not governed by rules are not necessarily random. Many thinkers have thought otherwise, because they supposed that rational actions and decisions were so because they were in conformity with some explicit or implicit rule. Even great creative artists have a style peculiarly their own, and it is natural to think that their style is characterized by some finite description, with some parameters left undetermined, and that random choices of these parameters is what produces different masterpiece of the artist's oeuvre. But this need not be so: the disjunction between the random variation and the finished specification is not exhaustive. For in the case of First-order Peano Arithmetic there are Gödelian formulae (many, in fact infinitely many, one for each system of coding) which are not assigned truth-values by the rules of the system, and which could therefore be assigned either TRUE or FALSE, each such assignment yielding a logically possible, consistent system. These systems are random vaunts, all satisfying the core description of Peano Arithmetic. But among them there is one, the one that assigns TRUE to all the Gödelian formulae which is reasonable, characterizing standard arithmetic, although not more in accordance with the specification of Peano Arithmetic than any of the others. So there is some sort of reasonableness, pocking out this one instantiation of the specification in preference to all the others which is reasonable and right, though not any more in accordance with the antecedently formulated rules than any other instantiation.

The critic may still complain that doing arithmetic in non-non-standard models is still a pretty boring activity; but that is not the point. Gödel's theorem shows that there is conceptual room for creativity, by allowing that to be reasonable is not necessarily to be in accordance with a

rule. We can see how it works out with the style of great creative artists. Titian, or Bach, or Shakespeare, develop a style which is peculiarly their own, and which we can learn to recognise. But it is not static. Up to a point they can produce works that are variations on a theme, but beyond that point we begin to criticize, and say that they are painting, composing, or writing, according to a formula; it is the mark of second-rate artists to be content to go on doing just that, but the genius is not content to rest on his laurels, but seeks to go further, and innovate, breaking out of the mould that his previous style was coming to constitute for him. Instead of there just being a formula to which all his work conformed, he produces work which differs in some significant respect from what he had been doing, and this difference is not just a random one, but one which ex post facto we recognise as essentially right, even though it was not required by the previous specification of his style. (19) Thus, though the Gödelian formula is not a very interesting formula to enunciate, the Gödelian argument argues strongly for creativity, first in ruling out any reductionist account of the mind that would show us to be, au fond, necessarily unoriginal automata, and secondly by proving that the conceptual space exists in which it intelligible to speak of someone's being creative, without having to hold that he must be either acting at random or else in accordance with an antecedently specifiable rule.

Notes

(*) A talk given on 25/5/96 at a BSPS conference in Oxford. For the copyright of the papers of J.R. Lucas see http://users.ox.ac.uk/~jrlucas/ back

(1) Journal of Behavioral and Brain Sciences, 13:4, 1990, p.693. back

(2) Kurt Gödel*: Collected Works, III*, ed. Feferman, Oxford, 1995, p. 310. back

(3) It is necessary to stress this point, as many critics try to distance Gödel's thought from that of Lucas and Penrose. Although there are significant differences between the three approaches, their conclusions are substantially similar. See H.Wang, *From Mathematics to Philosophy*, London, 1974, pp.324-326; and *Reflections on Kurt Gödel*, MIT Press, Cambridge, Mass., USA, 1987, p.48, pp.117-118. back

(4) Minds, Machines, and Gödel, Philosophy, 36, 1961, p.126; *The Modeling of Mind*, Kenneth M.Sayre and Frederick J.Crosson, eds., Notre Dame Press, 1963, pp.269-270; *Minds and Machines*, ed Alan Ross Anderson, Prentice-Hall, 1954, pp.58-59. back

(5) Thomas M.Breuer, Journal of Behavioral and Brain Sciences, 13:4, 1990, p.657; he continues:

Penrose's argument may be cautious first steps towards changing both of these facts, but I feel they are still much too tentative and informal to require serious reconsideration of the marriage of AI and the Turing model of computation. back

(6) See Hilary Putnam "Minds and Machines", in Sidney Hook, ed., *Dimensions of* Mind. A Symposium, New York, 1960; reprinted in A. R. Anderson, *Minds and Machines*, Prentice-Hall, 1964, pp. 72-97. back

(7) Chris Mortensen, Journal of Behavioral and Brain Sciences, 13:4, 1990, p.678. back

(8) George Boolos, Journal of Behavioral and Brain Sciences, 13:4, 1990, p.655:

I suggest that we do not know that we are not in the same situation vis-a-vis ZF that Frege was in with respect to naive set theory (or, more accurately, the system of his Basic Lavs of Arithmetic) before receiving, in June 1902, the famous letter from Russell, showing the derivability in his system of Russell's paradox.

Martin Davis, Journal of Behavioral and Brain Sciences, 13:4, 1990, p.660:

. . . convincing oneself that the given axioms are indeed consistent, since otherwise we will have no reason to believe that the Gödel sentence is true. But here things are quite murky: Great Iogicians (Frege, Curry, Church, Quine, Rosser) have managed to propose quite serious systems of logic which later have turned out to be inconsistent. back

(9) pp.120-124/263-268/52-56. back
(10) In addition to those noted in fn.8 above, it is worth citing:
1. S.Guccione, Journal of Behavioral and Brain Sciences, 16:3, 1993, p.612: *The most conclusive and immediate objection against this argument is Putnam's (1960) observation that (following Gödel's theorem) the human mind can only demonstrate the implication: "If S is consistent, then Gs is true.";*
2. Alexis Manaster-Ramer, Walter J. Savitch and Wlodek Zadrozny: *But all this - and more - depends on granting Penrose's argument, and this we should not do. The error is a small but lethal flaw in his presentation, and application, of Gödel's theorem. For Gödel does not say that a certain proposition P is true but unprovable in a formal system F, but merely that P is true but unprovable if F is consistent. Penrose notes that if F is inconsistent then P is provable but false, but then makes the inexplicable mistake of assuming that, "Our formal system should not be so badly constructed that it actually allows false propositions to be proved!" (pp. 107-108). Without this, only the conditional can be proved (and this can be done algorithmically!).*
3. G.Boolos, in Kurt Gödel: Collected Works, III, ed. Feferman, Oxford, 1995, pp.296, 295: *The classic reply to these views <of Ernest Nagel and James R. Newman (1958), J.R. Lucas (1961), and Roger Penrose (1989)> was given by Hilary Putnam (1960): Merely to find from a given machine M a statement S for which it can be proved that M, if consistent, cannot prove S is not to prove S---even if M is consistent. It is fair to say that the arguments of these writers have as yet obtained little credence.*
4. David J.Chalmers, Review of Shadows of the Mind, in PSYCHE: an interdisciplinary journal of research on consciousness, 2(9), June, 1995. Filename: psyche-95-2- 09-shadows-7-chalmers, (an elaboration of his review in Scientific American, June 1995, pp. 117-18. back
(11) Paul Benacerraf, God, the Devil and Gödel, The Monist, 1967, pp. 9-32. back
(12) Judson C. Webb, Mechanism, *Mentalism and Metamathematics; An Essay on Finitism*, Dordrecht, 1980, p.230. back
(13) I.J. Good, Human and Machine Logic, British Journal for the Philosophy of Science, 18, 1967, pp. 144-147; and "Gödel's Theorem is a Red Herring", British Journal for the Philosophy of Science, 19, 1968, pp. 357-8. back
(14) This point is conceded by Douglas R.Hofstadter, *Gödel, Escher, Bach*, New York, 1979, p.475, and R.W.Kentridge, Journal of Behavioral and Brain Sciences, 13:4, 1990, p.671, who seem to think it is a criticism, not a support, of my argument. back
(15) pp.117-118/262-262/49-50. back
(16) See R.W.Kentridge, Journal of Behavioral and Brain Sciences, 13:4, 1990, p.671: What it <this demonstration> actually shows is that we can do better than one particular algorithm H. It is easy to see how to construct a new algorithm H', a modification of H in which H'(k;k)=O, which does just as well as we do. Therefore, the example does not show that we think nonalgorithmically; all it shows is that we can prove something that one particular algorithm cannot. Adina Roskies, Journal of Behavioral and Brain Sciences, 13:4, 1990, p.682: Briefly, Penrose conflates the ability to solve an instance of a noncomputable problem with the ability to solve all instances of that problem. Only the latter would entail solving it nonalgorithmically. (my emphasis) back
(17) Thus J.Higginbotham, Journal of Behavioral and Brain Sciences, 13:4, 1990, p.668: Lucas's original argument to this effect is notoriously suggestive but vague, and critical literature on it has often taken the form of making the argument more precise, and then showing that in the precise form envisaged it fails to prove the case. (Penrose notes some of this literature, but does not discuss it directly in the text.) back
(18) See Crispin Wright, *Realism, Meaning and Truth*, 2nd ed., Blackwell, Oxford, 1993, p.351. back

(19) For further elucidation of the difference between there being a specification to which all cases belong and there being some case that differs from an antecedent specification, but is right to do so, see J.R.Lucas, ``The Lesbian Rule'', Philosophy, 1955, pp.195-213. back