Minds, Machines and Gödel

Author(s): J. R. Lucas

Source: *Philosophy*, Apr. - Jul., 1961, Vol. 36, No. 137 (Apr. - Jul., 1961), pp. 112-127

Published by: Cambridge University Press on behalf of Royal Institute of Philosophy

Stable URL: https://www.jstor.org/stable/3749270

MINDS, MACHINES AND GÖDEL[1]

J. R. LUCAS

GÖDEL'S Theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines. So also has it seemed to many other people: almost every mathematical logician I have put the matter to has confessed to similar thoughts, but has felt reluctant to commit himself definitely until he could see the whole argument set out, with all objections fully stated and properly met.[2] This I attempt to do.

Gödel's theorem states that in any consistent system which is strong enough to produce simple arithmetic there are formulae which cannot be proved-in-the-system, but which we can see to be true. Essentially, we consider the formula which says, in effect, "This formula is unprovable-in-the-system". If this formula were provable-in-the-system, we should have a contradiction: for if it were provable-in-the-system, then it would not be unprovable-in-the-system, so that "This formula is unprovable-in-the-system" would be false: equally, if it were provable-in-the-system, then it would not be false, but would be true, since in any consistent system nothing false can be proved-in-the-system, but only truths. So the formula "This formula is unprovable-in-the-system" is not provable-in-the-system, but unprovable-in-the-system. Further, if the formula "This formula is unprovable-in-the-system" is unprovable-in-the-system, then it is true that that formula is unprovable-in-the-system, that is, "This formula is unprovable-in-the-system" is true.

The foregoing argument is very fiddling, and difficult to grasp fully: it is helpful to put the argument the other way round, consider the possibility that "This formula is unprovable-in-the-system" might be false, show that that is impossible, and thus that the formula is true; whence it follows that it is unprovable. Even so, the argument remains persistently unconvincing: we feel that there must be a catch in it somewhere. The whole labour of Gödel's theorem is to show that there is no catch anywhere, and that the result can

[1] A paper read to the Oxford Philosophical Society on October 30, 1959.

[2] See A. M. Turing: "Computing Machinery and Intelligence": *Mind*, 1950, pp. 433–60, reprinted in *The World of Mathematics*, edited by James R. Newman, pp. 2099–123; and K. R. Popper: "Indeterminism in Quantum Physics and Classical Physics"; *British Journal for Philosophy of Science*, Vol. I (1951), pp. 179–88. The question is touched upon by Paul Rosenbloom; *Elements of Mathematical Logic*; pp. 207–8; Ernest Nagel and James R. Newman; *Gödel's proof*, pp. 100—2; and by Hartley Rogers; *Theory of Recursive Functions and Effective Computability* (mimeographed), 1957, Vol. I, pp. 152 ff.

be established by the most rigorous deduction; it holds for all formal systems which are (i) consistent, (ii) adequate for simple arithmetic —i.e. contain the natural numbers and the operations of addition and multiplication—and it shows that they are incomplete—i.e. contain unprovable, though perfectly meaningful, formulae, some of which, moreover, we, standing outside the system, can see to be true.

Gödel's theorem must apply to cybernetical machines, because it is of the essence of being a machine, that it should be a concrete instantiation of a formal system. It follows that given any machine which is consistent and capable of doing simple arithmetic, there is a formula which it is incapable of producing as being true—i.e. the formula is unprovable-in-the-system—but which we can see to be true. It follows that no machine can be a complete or adequate model of the mind, that minds are essentially different from machines.

We understand by a cybernetical machine an apparatus which performs a set of operations according to a definite set of rules. Normally we "programme" a machine: that is, we give it a set of instructions about what it is to do in each eventuality; and we feed in the initial "information" on which the machine is to perform its calculations. When we consider the possibility that the mind might be a cybernetical mechanism we have such a model in view; we suppose that the brain is composed of complicated neural circuits, and that the information fed in by the senses is "processed" and acted upon or stored for future use. If it is such a mechanism, then given the way in which it is programmed—the way in which it is "wired up"—and the information which has been fed into it, the response—the "output"—is determined, and could, granted sufficient time, be calculated. Our idea of a machine is just this, that its behaviour is completely determined by the way it is made and the incoming "stimuli": there is no possibility of its acting on its own: given a certain form of construction and a certain input of information, then it must act in a certain specific way. We, however, shall be concerned not with what a machine *must* do, but with what it *can* do. That is, instead of considering the whole set of rules which together determine exactly what a machine will do in given circumstances, we shall consider only an outline of those rules, which will delimit the possible responses of the machine, but not completely. The complete rules will determine the operations completely at every stage; at every stage there will be a definite instruction, e.g. "If the number is prime and greater than two add one and divide by two: if it is not prime, divide by its smallest factor": we, however, will consider the possibility of there being alternative instructions, e.g. "In a fraction you may divide top and bottom by *any* number which is a factor of both numerator and denominator". In thus

relaxing the specification of our model, so that it is no longer completely determinist, though still entirely mechanistic, we shall be able to take into account a feature often proposed for mechanical models of the mind, namely that they should contain a randomizing device. One could build a machine where the choice between a number of alternatives was settled by, say, the number of radium atoms to have disintegrated in a given container in the past half-minute. It is *prima facie* plausible that our brains should be liable to random effects: a cosmic ray might well be enough to trigger off a neural impulse. But clearly in a machine a randomizing device could not be introduced to choose any alternative whatsoever: it can only be permitted to choose between a number of allowable alternatives. It is all right to add *any* number chosen at random to both sides of an equation, but not to add one number to one side and another to the other. It is all right to choose to prove one theorem of Euclid rather than another, or to use one method rather than another, but not to "prove" something which is not true, or to use a "method of proof" which is not valid. Any randomizing devices must allow choices only between those operations which will not lead to inconsistency: which is exactly what the relaxed specification of our model specifies. Indeed, one might put it this way: instead of considering what a completely determined machine *must* do, we shall consider what a machine might be able to do if it had a randomizing device that acted whenever there were two or more operations possible, none of which could lead to inconsistency.

If such a machine were built to produce theorems about arithmetic (in many ways the simplest part of mathematics), it would have only a finite number of components, and so there would be only a finite number of types of operation it could do, and only a finite number of initial assumptions it could operate on. Indeed, we can go further, and say that there would only be a *definite* number of types of operation, and of initial assumptions, that could be built into it. Machines are definite: anything which was indefinite or infinite we should not count as a machine. Note that we say number of *types* of operation, not number of operations. Given sufficient time, and provided that it did not wear out, a machine could go on repeating an operation indefinitely: it is merely that there can be only a definite number of different *sorts* of operation it can perform.

If there are only a definite number of types of operation and initial assumptions built into the system, we can represent them all by suitable symbols written down on paper. We can parallel the operation by rules ("rules of inference" or "axiom schemata") allowing us to go from one or more formulae (or even from no formula at all) to another formula, and we can parallel the initial

assumptions (if any) by a set of initial formulae ("primitive propositions", "postulates" or "axioms"). Once we have represented these on paper, we can represent every single operation: all we need do is to give formulae representing the situation before and after the operation, and note which rule is being invoked. We can thus represent on paper any possible sequence of operations the machine might perform. However long the machine went on operating, we could, given enough time, paper and patience, write down an analogue of the machine's operations. This analogue would in fact be a formal proof: every operation of the machine is represented by the application of one of the rules: and the conditions which determine for the machine whether an operation can be performed in a certain situation, become, in our representation, conditions which settle whether a rule can be applied to a certain formula, i.e. formal conditions of applicability. Thus, construing our rules as rules of inference, we shall have a proof-sequence of formulae, each one being written down in virtue of some formal rule of inference having been applied to some previous formula or formulae (except, of course, for the initial formulae, which are given because they represent initial assumptions built into the system). The conclusions it is possible for the machine to produce as being true will therefore correspond to the theorems that can be proved in the corresponding formal system. We now construct a Gödelian formula in this formal system. This formula cannot be *proved-in-the-system*. Therefore the machine cannot produce the corresponding formula as being true. But *we* can see that the Gödelian formula is true: any rational being could follow Gödel's argument, and convince himself that the Gödelian formula, although unprovable-in-the-given-system, was nonetheless—in fact, for that very reason—true. Now any mechanical model of the mind must include a mechanism which can enunciate truths of arithmetic, because this is something which minds can do: in fact, it is easy to produce mechanical models which will in many respects produce truths of arithmetic far better than human beings can. But in this one respect they cannot do so well: in that for every machine there is a truth which it cannot produce as being true, but which a mind can. This shows that a machine cannot be a complete and adequate model of the mind. It cannot do *everything* that a mind can do, since however much it can do, there is always something which it cannot do, and a mind can. This is not to say that we cannot build a machine to simulate *any* desired piece of mind-like behaviour: it is only that we cannot build a machine to simulate *every* piece of mind-like behaviour. We can (or shall be able to one day) build machines capable of reproducing bits of mind-like behaviour, and indeed of outdoing the performances of human minds: but however good the machine is, and however much better

it can do in nearly all respects than a human mind can, it always has this one weakness, this one thing which it cannot do, whereas a mind can. The Gödelian formula is the Achilles' heel of the cybernetical machine. And therefore we cannot hope ever to produce a machine that will be able to do all that a mind can do: we can never, not even in principle, have a mechanical model of the mind.

This conclusion will be highly suspect to some people. They will object first that we cannot have it both that a machine *can* simulate *any* piece of mind-like behaviour, and that it *cannot* simulate *every* piece. To some it is a contradiction: to them it is enough to point out that there is no contradiction between the fact that for any natural number there can be produced a greater number, and the fact that a number cannot be produced greater than every number. We can use the same analogy also against those who, finding a formula their first machine cannot produce as being true, concede that that machine is indeed inadequate, but thereupon seek to construct a second, more adequate, machine, in which the formula *can* be produced as being true. This they can indeed do: but then the second machine will have a Gödelian formula all of its own, constructed by applying Gödel's procedure to the formal system which represents its (the second machine's) own, enlarged, scheme of operations. And this formula the second machine will not be able to produce as being true, while a mind will be able to see that it is true. And if now a third machine is constructed, able to do what the second machine was unable to do, exactly the same will happen: there will be yet a third formula, the Gödelian formula for the formal system corresponding to the third machine's scheme of operations, which the third machine is unable to produce as being true, while a mind will still be able to see that it is true. And so it will go on. However complicated a machine we construct, it will, if it is a machine, correspond to a formal system, which in turn will be liable to the Gödel procedure for finding a formula unprovable-in-that-system. This formula the machine will be unable to produce as being true, although a mind can see that it is true. And so the machine will still not be an adequate model of the mind. We are trying to produce a model of the mind which is mechanical— which is essentially "dead"—but the mind, being in fact "alive", can always go one better than any formal, ossified, dead, system can. Thanks to Gödel's theorem, the mind always has the last word.

A second objection will now be made. The procedure whereby the Gödelian formula is constructed is a standard procedure—only so could we be sure that a Gödelian formula can be constructed for every formal system. But if it is a standard procedure, then a machine should be able to be programmed to carry it out too. We could construct a machine with the usual operations, and in addition an

operation of going through the Gödel procedure, and then producing the conclusion of that procedure as being true; and then repeating the procedure, and so on, as often as required. This would correspond to having a system with an additional rule of inference which allowed one to add, as a theorem, the Gödelian formula of the rest of the formal system, and then the Gödelian formula of this new, strengthened formal system, and so on. It would be tantamount to adding to the original formal system an infinite sequence of axioms, each the Gödelian formula of the system hitherto obtained. Yet even so, the matter is not settled: for the machine with a Gödelizing operator, as we might call it, is a *different* machine from the machines without such an operator; and, although the machine with the operator would be able to do those things in which the machines without the operator were outclassed by a mind, yet we might expect a mind, faced with a machine that possessed a Gödelizing operator, to take this into account, and out-Gödel the new machine, Gödelizing operator and all. This has, in fact, proved to be the case. Even if we adjoin to a formal system the infinite set of axioms consisting of the successive Gödelian formulae, the resulting system is still incomplete, and contains a formula which cannot be proved-in-the-system, although a rational being can, standing outside the system, see that it is true.[1] We had expected this, for even if an infinite set of axioms were added, they would have to be specified by some finite rule or specification, and this further rule or specification could then be taken into account by a mind considering the enlarged formal system. In a sense, just because the mind has the last word, it can always pick a hole in any formal system presented to it as a model of its own workings. The mechanical model must be, in some sense, finite and definite: and then the mind can always go one better.

This is the answer to one objection put forward by Turing.[2] He argues that the limitation to the powers of a machine do not amount to anything much. Although each individual machine is incapable of getting the right answer to some questions, after all each individual human being is fallible also: and in any case "our superiority can only be felt on such an occasion in relation to the one machine over which we have scored our petty triumph. There would be no question of triumphing simultaneously over *all* machines". But this is not the point. We are not discussing whether machines or minds are superior, but whether they are the same. In some respect machines are undoubtedly superior to human minds; and the question on which they are stumped is admittedly, a rather niggling, even

[1] Gödel's original proof applies; *v.* § 1 init. § 6 init. of his Lectures at the Institute of Advanced Study, Princeton, N.J., U.S.A., 1934.

[2] *Mind*, 1950, pp. 444–5; Newman, p. 2110.

trivial, question. But it is enough, enough to show that the machine is *not the same* as a mind. True, the machine can do many things that a human mind cannot do: but if there is of necessity something that the machine cannot do, though the mind can, then, however trivial the matter is, we cannot equate the two, and cannot hope ever to have a mechanical model that will adequately represent the mind. Nor does it signify that it is only an individual machine we have triumphed over: for the triumph is not over only *an* individual machine, but over *any* individual that anybody cares to specify—in Latin *quivis* or *quilibet*. not *quidam*—and a mechanical model of a mind must be an individual machine. Although it is true that any particular "triumph" of a mind over a machine could be "trumped" by another machine able to produce the answer the first machine could not produce, so that "there is no question of triumphing simultaneously over all machines", yet this is irrelevant. What is at issue is not the unequal contest between one mind and all machines, but whether there could be any, single, machine that could do all a mind can do. For the mechanist thesis to hold water, it must be possible, in principle, to produce a model, a single model, which can do everything the mind can do. It is like a game.[1] The mechanist has first turn. He produces *a—any*, but only a *definite one* —mechanical model of the mind. I point to something that it cannot do, but the mind can. The mechanist is free to modify his example, but each time he does so, I am entitled to look for defects in the revised model. If the mechanist can devise a model that I cannot find fault with, his thesis is established: if he cannot, then it is not proven: and since—as it turns out—he necessarily cannot, it is refuted. To succeed, he must be able to produce some definite mechanical model of the mind—any one he likes, but one he can specify, and will stick to. But since he cannot, in principle cannot, produce any mechanical model that is adequate, even though the point of failure is a minor one, he is bound to fail, and mechanism must be false.

Deeper objections can still be made. Gödel's theorem applies to deductive systems, and human beings are not confined to making only deductive inferences. Gödel's theorem applies only to consistent systems, and one may have doubts about how far it is permissible to assume that human beings are consistent. Gödel's theorem applies only to formal systems, and there is no *a priori* bound to human ingenuity which rules out the possibility of our contriving some replica of humanity which was not representable by a formal system.

[1] For a similar type of argument, see J. R. Lucas: "The Lesbian Rule"; PHILOSOPHY, July 1955, pp. 202–6; and "On not worshipping Facts"; *The Philosophical Quarterly*, April 1958, p. 144.

Human beings are not confined to making deductive inferences, and it has been urged by C. G. Hempel[1] and Hartley Rogers[2] that a fair model of the mind would have to allow for the possibility of making non-deductive inferences, and these might provide a way of escaping the Gödel result. Hartley Rogers makes the specific suggestion that the machine should be programmed to entertain various propositions which had not been proved or disproved, and on occasion to add them to its list of axioms. Fermat's last theorem or Goldbach's conjecture might thus be added. If subsequently their inclusion was found to lead to a contradiction, they would be dropped again, and indeed in those circumstances their negations would be added to the list of theorems. In this sort of way a machine might well be constructed which was able to produce as true certain formulae which could not be proved from its axioms according to its rules of inference. And therefore the method of demonstrating the mind's superiority over the machine might no longer work.

The construction of such a machine, however, presents difficulties. It cannot accept all unprovable formulae, and add them to its axioms, or it will find itself accepting both the Gödelian formula and its negation, and so be inconsistent. Nor would it do if it accepted the first of each pair of undecidable formulae, and, having added that to its axioms, would no longer regard its negation as undecidable, and so would never accept it too: for it might happen on the wrong member of the pair: it might accept the negation of the Gödelian formula rather than the Gödelian formula itself. And the system constituted by a normal set of axioms with the negation of the Gödelian formula adjoined, although not inconsistent, is an unsound system, not admitting of the natural interpretation. It is something like non-Desarguian geometries in two dimensions: not actually inconsistent, but rather wrong, sufficiently much so to disqualify it from serious consideration. A machine which was liable to infelicities of that kind would be no model for the human mind.

It becomes clear that rather careful criteria of selection of unprovable formulae will be needed. Hartley Rogers suggests some possible ones. But once we have rules generating new axioms, even if the axioms generated are only provisionally accepted, and are liable to be dropped again if they are found to lead to inconsistency, then we can set about doing a Gödel on this system, as on any other. We are in the same case as when we had a rule generating the infinite set of Gödelian formulae as axioms. In short, however a machine is designed, it must proceed either at random or according to definite rules. In so far as its procedure is random, we cannot outsmart it:

[1] In private conversation.

[2] *Theory of Recursive Functions and Effective Computability*, 1957, Vol. I, pp. 152 ff.

but its performance is not going to be a convincing parody of intelligent behaviour: in so far as its procedure is in accordance with definite rules, the Gödel method can be used to produce a formula which the machine, according to those rules, cannot assert as true, although we, standing outside the system, can see it to be true.[1]

Gödel's theorem applies only to consistent systems. All that we can prove *formally* is that *if* the system is complete, then the Gödelian formula is unprovable-in-the-system. To be able to say categorically that the Gödelian formula is unprovable-in-the-system, and therefore true, we must not only be dealing with a consistent system, but be able to say that it is consistent. And, as Gödel showed in his second theorem—a corollary of his first—it is impossible to prove in a consistent system that that system is consistent. Thus in order to fault the machine by producing a formula of which we can say both that it is true and that the machine cannot produce it as true, we have to be able to say that the machine (or, rather, its corresponding formal system) is consistent; and there is no absolute proof of this. All we can do is to examine the machine and see if it appears consistent. There always remains the possibility of some inconsistency not yet detected. At best we can say that the machine is consistent, provided we are. But by what right can we do this? Gödel's second theorem seems to show that a man cannot assert his own consistency, and so Hartley Rogers[2] argues that we cannot really use Gödel's first theorem to counter the mechanist thesis unless we can say that "there are distinctive attributes which enable a human being to transcend this last limitation and assert his own consistency while still remaining consistent".

A man's untutored reaction if his consistency is questioned is to affirm it vehemently: but this, in view of Gödel's second theorem, is taken by some philosophers as evidence of his actual inconsistency. Professor Putnam[3] has suggested that human beings are machines, but inconsistent machines. If a machine were wired to correspond to an inconsistent system, then there would be no well-formed formula which it could not produce as true; and so in no way could it be proved to be inferior to a human being. Nor could we make its inconsistency a reproach to it—are not men inconsistent too? Certainly women are, and politicians; and even male non-politicians

[1] Gödel's original proof applies if the rule is such as to generate a primitive recursive class of additional formulae; *v.* § 1 init. and § 6 init. of his Lectures at the Institute of Advanced Study, Princeton, N.J., U.S.A., 1934. It is in fact sufficient that the class be recursively enumerable. *v.* Barkley Rosser: "Extensions of some theorems of Gödel and Church", *Journal of Symbolic Logic*, Vol. I, 1936, pp. 87–91.

[2] Op. cit., p. 154.

[3] University of Princeton, N.J., U.S.A. in private conversation.

contradict themselves sometimes, and a single inconsistency is enough to make a system inconsistent.

The fact that we are all sometimes inconsistent cannot be gainsaid, but from this it does not follow that we are tantamount to inconsistent systems. Our inconsistencies are mistakes rather than set policies. They correspond to the occasional malfunctioning of a machine, not its normal scheme of operations. Witness to this that we eschew inconsistencies when we recognize them for what they are. If we really were inconsistent machines, we should remain content with our inconsistencies, and would happily affirm both halves of a contradiction. Moreover, we would be prepared to say absolutely anything—which we are not. It is easily shown[1] that in an inconsistent formal system everything is provable, and the requirement of consistency turns out to be just that not everything can be proved in it—it is not the case that "anything goes". This surely is a characteristic of the mental operations of human beings: they are selective: they do discriminate between favoured—true—and unfavoured—false—statements: when a person is prepared to say anything, and is prepared to contradict himself without any qualm or repugnance, then he is adjudged to have "lost his mind". Human beings, although not perfectly consistent, are not so much inconsistent as fallible.

A fallible but self-correcting machine would still be subject to Gödel's results. Only a fundamentally inconsistent machine would escape. Could we have a fundamentally inconsistent, but at the same time self-correcting machine, which both would be free of Gödel's results and yet would not be trivial and entirely unlike a human being? A machine with a rather *recherché* inconsistency wired into it, so that for all normal purposes it was consistent, but when presented with the Gödelian sentence was able to prove it?

There are all sorts of ways in which undesirable proofs might be obviated. We might have a rule that whenever we have proved p and not-p, we examine their proofs and reject the longer. Or we might arrange the axioms and rules of inference in a certain order, and when a proof leading to an inconsistency is proffered, see what axioms and rules are required for it, and reject that axiom or rule which comes last in the ordering. In some such way as this we could have an inconsistent system, with a stop-rule, so that the inconsistency was never allowed to come out in the form of an inconsistent formula.

The suggestion at first sight seems attractive: yet there is something deeply wrong. Even though we might preserve the façade of consistency by having a rule that whenever two inconsistent formulae

[1] See, e.g., Alonzo Church: *Introduction to Mathematical Logic*, Princeton, Vol. I, § 17, p. 108.

appear we were to reject the one with the longer proof, yet such a rule would be repugnant in our logical sense. Even the less arbitrary suggestions are too arbitrary. No longer does the system operate with certain definite rules of inference on certain definite formulae. Instead, the rules apply, the axioms are true, provided . . . we do not happen to find it inconvenient. We no longer know where we stand. One application of the rule of Modus Ponens may be accepted while another is rejected: on one occasion an axiom may be true, on another apparently false. The system will have ceased to be a formal logical system, and the machine will barely qualify for the title of a model for the mind. For it will be far from resembling the mind in its operations: the mind does indeed try out dubious axioms and rules of inference; but if they are found to lead to contradiction, they are rejected altogether. We try out axioms and rules of inference provisionally—true: but we do not keep them, once they are found to lead to contradictions. We may seek to replace them with others, we may feel that our formalization is at fault, and that though some axiom or rule of inference of this sort is required, we have not been able to formulate it quite correctly: but we do not retain the faulty formulations without modification, merely with the proviso that when the argument leads to a contradiction we refuse to follow it. To do this would be utterly irrational. We should be in the position that on some occasions when supplied with the premisses of a Modus Ponens, say, we applied the rule and allowed the conclusion, and on other occasions we refused to apply the rule, and disallowed the conclusion. A person, or a machine, which did this without being able to give a good reason for so doing, would be accounted arbitrary and irrational. It is part of the concept of "arguments" or "reasons" that they are in some sense general and universal: that if Modus Ponens is a valid method of arguing when I am establishing a desired conclusion, it is a valid method also when you, my opponent, are establishing a conclusion I do not want to accept. We cannot pick and choose the times when a form of argument is to be valid; not if we are to be reasonable. It is of course true, that with our informal arguments, which are not fully formalized, we do distinguish between arguments which are at first sight similar, adding further reasons why they are nonetheless not really similar: and it might be maintained that a machine might likewise be entitled to distinguish between arguments at first sight similar, if it had good reason for doing so. And it might further be maintained that the machine had good reason for rejecting those patterns of argument it did reject, indeed the best of reasons, namely the avoidance of contradiction. But that, if it is a reason at all, is too good a reason. We do not lay it to a man's credit that he avoids contradiction merely by refusing to accept those arguments which would lead him to it, for no other

reason than that otherwise he would be led to it. Special pleading rather than sound argument is the name for that type of reasoning. No credit accrues to a man who, clever enough to see a few moves of argument ahead, avoids being brought to acknowledge his own inconsistency, by stonewalling as soon as he sees where the argument will end. Rather, we account him inconsistent too, not, in his case, because he affirmed and denied the same proposition, but because he used and refused to use the same rule of inference. A stop-rule on actually enunciating an inconsistency is not enough to save an inconsistent machine from being called inconsistent.

The possibility yet remains that we are inconsistent, and there is no stop-rule, but the inconsistency is so *recherché* that it has never turned up. After all, *naïve* set-theory, which was deeply embedded in common-sense ways of thinking did not turn out to be inconsistent. Can we be sure that a similar fate is not in store for simple arithmetic too? In a sense we cannot, in spite of our great feeling of certitude that our system of whole numbers which can be added and multiplied together is never going to prove inconsistent. It is just conceivable we might find we had formalized it incorrectly. If we had, we should try and formulate anew our intuitive concept of number, as we have our intuitive concept of a set. If we did this, we should of course recast our system: our present axioms and rules of inference would be utterly rejected: there would be no question of our using and not using them in an "inconsistent" fashion. We should, once we had recast the system, be in the same position as we are now, possessed of a system believed to be consistent, but not provably so. But then could there not be some other inconsistency? It is indeed a possibility. But again no inconsistency once detected will be tolerated. We are determined not to be inconsistent, and are resolved to root out inconsistency, should any appear. Thus, although we can never be completely certain or completely free of the risk of having to think out our mathematics again, the ultimate position must be one of two: either we have a system of simple arithmetic which to the best of our knowledge and belief is consistent: or there is no such system possible. In the former case we are in the same position as at present: in the latter, if we find that no system containing simple arithmetic can be free of contradictions, we shall have to abandon not merely the whole of mathematics and the mathematical sciences, but the whole of thought.

It may still be maintained that although a man must in this sense assume, he cannot properly affirm, his own consistency without thereby belying his words. We may be consistent; indeed we have every reason to hope that we are: but a necessary modesty forbids us from saying so. Yet this is not quite what Gödel's second theorem states. Gödel has shown that in a consistent system a formula

stating the consistency of the system cannot be proved *in that system*. It follows that a machine, if consistent, cannot produce as true an assertion of its own consistency: hence also that a mind, *if it were really a machine*, could not reach the conclusion that it was a consistent one. For a mind which is not a machine no such conclusion follows. All that Gödel has proved is that a mind cannot produce a formal proof of the consistency of a formal system inside the system itself: but there is no objection to going outside the system and no objection to producing informal arguments for the consistency either of a formal system or of something less formal and less systematized. Such informal arguments will not be able to be completely formalized: but then the whole tenor of Gödel's results is that we ought not to ask, and cannot obtain, complete formalization. And although it would have been nice if we could have obtained them, since completely formalized arguments are more coercive than informal ones, yet since we cannot have all our arguments cast into that form, we must not hold it against informal arguments that they are informal or regard them all as utterly worthless. It therefore seems to me both proper and reasonable for a mind to assert its own consistency: proper, because although machines, as we might have expected, are unable to reflect fully on their own performance and powers, yet to be able to be self-conscious in this way is just what we expect of minds: and reasonable, for the reasons given. Not only can we fairly say simply that we *know* we are consistent, apart from our mistakes, but we must in any case *assume* that we are, if thought is to be possible at all; moreover we are selective, we will not, as inconsistent machines would, say anything and everything whatsoever: and finally we can, in a sense, *decide* to be consistent, in the sense that we can resolve not to tolerate inconsistencies in our thinking and speaking, and to eliminate them, if ever they should appear, by withdrawing and cancelling one limb of the contradiction.

We can see how we might almost have expected Gödel's theorem to distinguish self-conscious beings from inanimate objects. The essence of the Gödelian formula is that it is self-referring. It says that "This formula is unprovable-in-this-system". When carried over to a machine, the formula is specified in terms which depend on the particular machine in question. The machine is being asked a question about its own processes. We are asking it to be self-conscious, and say what things it can and cannot do. Such questions notoriously lead to paradox. At one's first and simplest attempts to philosophize, one becomes entangled in questions of whether when one knows something one knows that one knows it, and what, when one is thinking of oneself, is being thought about, and what is doing the thinking. After one has been puzzled and bruised by this

problem for a long time, one learns not to press these questions: the concept of a conscious being is, implicitly, realized to be different from that of an unconscious object. In saying that a conscious being knows something, we are saying not only that he knows it, but that he knows that he knows it, and that he knows that he knows that he knows it, and so on, as long as we care to pose the question: there is, we recognize, an infinity here, but it is not an infinite regress in the bad sense, for it is the questions that peter out, as being pointless, rather than the answers. The questions are felt to be pointless because the concept contains within itself the idea of being able to go on answering such questions indefinitely. Although conscious beings have the power of going on, we do not wish to exhibit this simply as a succession of tasks they are able to perform, nor do we see the mind as an infinite sequence of selves and super-selves and super-super-selves. Rather, we insist that a conscious being is a unity, and though we talk about parts of the mind, we do so only as a metaphor, and will not allow it to be taken literally.

The paradoxes of consciousness arise because a conscious being can be aware of itself, as well as of other things, and yet cannot really be construed as being divisible into parts. It means that a conscious being can deal with Gödelian questions in a way in which a machine cannot, because a conscious being can both consider itself and its performance and yet not be other than that which did the performance. A machine can be made in a manner of speaking to "consider" its own performance, but it cannot take this "into account" without thereby becoming a different machine, namely the old machine with a "new part" added. But it is inherent in our idea of a conscious mind that it can reflect upon itself and criticize its own performances, and no extra part is required to do this: it is already complete, and has no Achilles' heel.

The thesis thus begins to become more a matter of conceptual analysis than mathematical discovery. This is borne out by considering another argument put forward by Turing.[1] So far, we have constructed only fairly simple and predictable artefacts. When we increase the complexity of our machines there may, perhaps, be surprises in store for us. He draws a parallel with a fission pile. Below a certain "critical" size, nothing much happens: but above the critical size, the sparks begin to fly. So too, perhaps, with brains and machines. Most brains and all machines are, at present, "sub-critical"—they react to incoming stimuli in a stodgy and uninteresting way, have no ideas of their own, can produce only stock responses —but a few brains at present, and possibly some machines in the future, are super-critical, and scintillate on their own account.

[1] *Mind*, 1950, p. 454; Newman, p. 2117–18.

Turing is suggesting that it is only a matter of complexity, and that above a certain level of complexity a qualitative difference appears, so that "super-critical" machines will be quite unlike the simple ones hitherto envisaged.

This may be so. Complexity often does introduce qualitative differences. Although it sounds implausible, it might turn out that above a certain level of complexity, a machine ceased to be predictable, even in principle, and started doing things on its own account, or, to use a very revealing phrase, it might begin to have a mind of its own. It might begin to have a mind of its own. It would begin to have a mind of its own when it was no longer entirely predictable and entirely docile, but was capable of doing things which we recognized as intelligent, and not just mistakes or random shots, but which we had not programmed into it. But then it would cease to be a machine, within the meaning of the act. What is at stake in the mechanist debate is not how minds are, or might be, brought into being, but how they operate. It is essential for the mechanist thesis that the mechanical model of the mind shall operate according to "mechanical principles", that is, that we can understand the operation of the whole in terms of the operations of its parts, and the operation of each part either shall be determined by its initial state and the construction of the machine, or shall be a random choice between a determinate number of determinate operations. If the mechanist produces a machine which is so complicated that this ceases to hold good of it, then it is no longer a machine for the purposes of our discussion, no matter how it was constructed. We should say, rather, that he had created a mind, in the same sort of sense as we procreate people at present. There would then be two ways of bringing new minds into the world, the traditional way, by begetting children born of women, and a new way by constructing very, very complicated systems of, say, valves and relays. When talking of the second way, we should take care to stress that although what was created looked like a machine, it was not one really, because it was not just the total of its parts. One could not tell what it was going to do merely by knowing the way in which it was built up and the initial state of its parts: one could not even tell the limits of what it could do, for even when presented with a Gödel-type question, it got the answer right. In fact we should say briefly that any system which was not floored by the Gödel question was *eo ipso* not a Turing machine, i.e. not a machine within the meaning of the act.

If the proof of the falsity of mechanism is valid, it is of the greatest consequence for the whole of philosophy. Since the time of Newton, the bogey of mechanist determinism has obsessed philosophers. If we were to be scientific, it seemed that we must look on human beings as

determined automata, and not as autonomous moral agents; if we were to be moral, it seemed that we must deny science its due, set an arbitrary limit to its progress in understanding human neurophysiology, and take refuge in obscurantist mysticism. Not even Kant could resolve the tension between the two standpoints. But now, though many arguments against human freedom still remain, the argument from mechanism, perhaps the most compelling argument of them all, has lost its power. No longer on this count will it be incumbent on the natural philosopher to deny freedom in the name of science: no longer will the moralist feel the urge to abolish knowledge to make room for faith. We can even begin to see how there could be room for morality, without its being necessary to abolish or even to circumscribe the province of science. Our argument has set no limits to scientific enquiry: it will still be possible to investigate the working of the brain. It will still be possible to produce mechanical models of the mind. Only, now we can see that no mechanical model will be completely adequate, nor any explanations in purely mechanist terms. We can produce models and explanations, and they will be illuminating: but, however far they go, there will always remain more to be said. There is no arbitrary bound to scientific enquiry: but no scientific enquiry can ever exhaust the infinite variety of the human mind.

Merton College, Oxford.