Use and Misuse of Gödel's Theorem

Shingo FUJITA*

Wang (1974) reports that in Gödel's opinion the two most interesting results proved with rigor about minds and machines are these :

- (1) The human mind is incapable of formulating (or mechanizing) all its mathematical intuitions. I.e. : if it has succeeded in formulating some of them, this very fact yields new intuitive knowledge, e.g. the consistency of this formalism. This fact may be called the 'incompleteness' of mathematics. On the other hand, on the basis of what has been proved so far, it remains possible that there may exist (and even be empirically discoverable) a theorem-proving machine which in fact is equivalent to mathematical intuition, but cannot be *proved* to be so, nor even be proved to yield only *correct* theorems of finitary number theory.
- (2) Either the human mind surpasses all machines (to be more precise : it can decide more number-theoretical questions than any machine) or else there exist number-theoretical questions undecidable for the human mind. (p. 324)

In the following I will call these two results "G(1)" and "G(2)" respectively.

Some of the questions which will be answered in this paper are these :

Question 1 : Why is it that there are *two* results, G(1) and G(2), which were proved with rigor ? What is the difference between G(1) and G(2) ?

Question 2 : Which is it that Gödel's theorem are to be applied to, minds or brains ? In other words, which do we intend a Turing machine to be a model of ?

Question 3 : Each of the results G(1) and G(2) have the form of disjunction. Which alternative should we choose ?

1. Gödel

Gödel (*1951) mentions G(2) as "the mathematically established fact" which seemed to him "of a great philosophical interest". It was so because of his philosophical standpoint, Platonism. G(1) was too weak to draw Platonism from his incompleteness theorems as one of their philosophical implications. Before going into his philosophical standpoint, we must look into his argument which led to the alleged rigorously established facts.

He starts with reformulating the first incompleteness theorem as a result about

* University of Tsukuba

a number-theoretical problem known as the diophantine problem (the explanation of which I omit here). The result says that “*whatever well-defined system of axioms and rules of inference may be chosen, there always exist diophantine problems of the type described which are undecidable by these axioms and rules, provided only that no false propositions of this type are derivable*”. (p. 308) Next comes the second incompleteness theorem, which says that “*for any well-defined system of axioms and rules, in particular, the proposition stating their consistency (or rather the equivalent number-theoretical propositions) is undemonstrable from these axioms and rules, provided these axioms and rules are consistent and suffice to derive a certain portion of finitistic arithmetic of integers*”. (pp. 308-9)

He then goes on to distinguish between “objective mathematics” and “subjective mathematics”. By “objective mathematics” he means the system of all true mathematical propositions, and by “subjective mathematics” the system of all demonstrable mathematical propositions. As far as objective mathematics is concerned, it follows from the second incompleteness theorem that no well-defined system of correct axioms can contain all of objective mathematics, because the proposition which states the consistency of the system cannot be proved in it, though true.

But what about subjective mathematics? Can we assert that all of subjective mathematics is contained in the formal system which we believe is sound? Actually Gödel’s answer to this question comprises his case against conventionalism and mechanism. What is at stake here is this: Even on the conventionalist view of logic and mathematics, the claim to proving the consistency of the system has to be met. We know, however, that the claim cannot be met according to his second incompleteness theorem. Gödel argues as follows. Assume that there should exist a certain well-defined system of axioms and rules and that we assert that the system could contain all demonstrable mathematical propositions. Then it follows that our assertion cannot be true. Because if we assert that all propositions it produces are correct (therefore consistent), then, by Gödel’s second theorem, it follows that the assertion is not derivable from the axioms and rules. Gödel says:

As to subjective mathematics, it is not precluded that there should exist a finite rule producing all its evident axioms. However, if such a rule exists, we with our human understanding could certainly never know it to be such, that is, we could never know with mathematical certainty that all propositions it produces are correct; or in other terms, we could perceive to be true only one proposition after the other, for any finite number of them. The assertion, however, that they are all true could at most be known with empirical certainty on the basis of a sufficient number of instances or by other inductive inferences. If it were so, this would mean that the human mind (in the realm of pure mathematics)

is equivalent to a finite machine that, however, is unable to understand completely its own functioning. (pp. 309-10)

Two points are worthy of remark. The first relates to the equivalence of the human mind to a finite machine which cannot be *proved* to be so. (To replace "the human mind" by "mathematical intuition", as was actually done by Wang (1974), is misleading, because the latter is imbued with some kind of philosophical standpoint.¹) In so far as subjective mathematics is concerned, there is no obstacle to supposing that "the human mind is equivalent to a finite machine, which however is unable to understand completely its own functioning". The reason is that to understand completely its own functioning means to know with mathematical certainty that all propositions it produces are correct (in contradistinction to saying that all the consequences from the axioms *one after the other* are correct) and therefore consistent, contrary to the second incompleteness theorem.

The second concerns "a finite machine which is equivalent to the human mind, but which cannot even be proved to yield only *correct* theorems of finitary number theory". As is well known, Gödel's theorem states that *if the formal system is consistent*, then there always exist such and such a formula.....etc. The hypothesis of *consistency* is stated as "*provided only that no false propositions of this type are derivable*" in the reformulation mentioned above. If it is the case that the human mind is equivalent to a finite machine and that the finite machine can be proved to yield only correct theorems (therefore, it can be proved to be consistent), then the hypothesis of *consistency* would not hold. This is the reason why the finite machine cannot even be proved to yield only correct theorems.

Gödel concludes, "The assertion that they are all true could at most be known with empirical certainty." "This would mean", he argues directly from the mere "empirical certainty" of the consistency of the system, "that the human mind (in the realm of pure mathematics) is equivalent to a finite machine that, however, is unable to understand completely its own functioning." (p. 310) This is another way of explaining the previous provisos. From the fact that "the consistency could at most be known with empirical certainty", it follows that there may exist a finite machine which in fact is equivalent to the human mind, but cannot be *proved* to be so [because no certainty can be given to the consistency beyond an empirical one], nor even be proved to yield only *correct* theorems of finitary number theory [because the claim that a finite machine can prove only *correct* theorems has at most empirical certainty]. This is the second alternative of G(1) mentioned at the outset.

Now I want to discuss G(2). The most important point to be stressed is that G(2), in contradistinction to G(1), resorts to the possible existence of absolutely unsolvable number-theoretic questions which would be undecidable even by the human mind. In his wording, G(2) runs as follows :

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 power of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified (where the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives). (p. 310)

Wang (1974) says, “Gödel thinks Hilbert was right in rejecting the second alternative.” It looks as if he is suggesting that Gödel did actually reject the second alternative of G(2). He says, “If it [the second alternative of G(2)] were true it would mean that human reason is utterly irrational by asking questions it cannot answer, while asserting emphatically that only reason can answer them”. (pp. 324–25) However, it should be noted that Gödel’s conclusion is drawn on the assumption that *the human mind is equivalent to a finite machine*. I will quote the passage in question in full.

This inability [of the human mind that is equivalent to a finite machine] to understand himself would then wrongly appear to him as its boundlessness or inexhaustibility. But, please, note that if it were so, this would in no way derogate from the incompleteness of objective mathematics. For if the human mind were equivalent to a finite machine, then objective mathematics not only would be incompletable in the sense of not being contained in any well-defined axiomatic system, but moreover there would exist *absolutely* unsolvable diophantine problems of the type described above, where the epithet “absolutely” means that they would be undecidable, not just within some particular axiomatic system, but by any mathematical proof the human mind can conceive. (p. 310)

This passage is extremely difficult to understand. On the one hand, “the human mind” that can conceive of mathematical proofs must be the human mind which is supposed to be equivalent to a finite machine, for otherwise the passage would be a *non sequitur*. But if the human mind were equivalent to a finite machine, then the fact that Gödel sides with Hilbert’s “rationalistic optimism” would have no point. On the other hand, if there were supposed to exist unsolvable diophantine problems which would be undecidable by any mathematical proof the human mind *per se* (i.e., the human mind as a human mind, in contrast to a finite machine) could conceive, this supposition would be a mere fantasy. In either case the argument is unjustified. I wonder if the passage is merely a Platonist statement which has the effect of making us believe that it is a logical inference.²

Another interpretation that suggest itself to me is, contrary to Wang’s (1974)

suggestion, is that $G(2)$ does not differ from $G(1)$, and that what Gödel rejected for a “rationalistic optimism” is the second alternative, not only of $G(2)$, but also of $G(1)$ as well. On this interpretation, the human mind which is unable to understand itself is supposed to be equivalent to an absolutely unsolvable number-theoretic question. One difficulty with this interpretation is that it makes no distinction between objective and subjective mathematics. Another difficulty is how a finite machine (or the human mind equivalent to it) could ask that absolutely unsolvable question. If it could not ask the question, then the second alternative would not be rejected for the sake of a “rationalistic optimism”. In any way the following extract from Wang (1996) is an evidence for this interpretation.

6.1.8. My incompleteness theorem makes it likely that mind is not mechanical, or else mind cannot understand its own mechanism. If my result is taken together with the rationalistic attitude which Hilbert had and which was not refuted by my results, then [we can infer] the sharp result that mind is not mechanical. This is so, because, if the mind were a machine, there would, contrary to this rationalistic attitude, exist number-theoretic questions undecidable for the human mind. (pp. 186-7)

Whichever interpretation one may adopt, the fact remains that $G(2)$ presupposes objective mathematics. And the fact that $G(2)$ presupposes objective mathematics explains clearly why Gödel preferred $G(2)$ to $G(1)$. While $G(1)$ represents the negative side of his theorem in the sense that it is concerned just with his criticism of mechanism and conventionalism, $G(2)$ has the positive effect of revealing the Platonistic aspect of mathematics. (Answer to our Question 1). If the human mind surpasses any finite machine, then, Gödel says, “the working of the mind cannot be reduced to the working of the brain.So apparently one is driven to take some vitalistic viewpoint”. (p. 311) But $G(2)$ is not justified. For the second alternative of $G(2)$ is not a derivation from his theorem, but a *petitio principii*.

What Gödel wanted to draw was that under either alternative of $G(2)$, “they are very decidedly opposed to materialistic philosophy”. He could indeed identify the brain with a finite machine. So far as the brain is concerned, he could concede to materialism as much as it would require. Even if in the development of the brain sciences it might finally turn out that “the brain suffices for the explanation of all mental phenomena and is a machine in the sense of Turing”, it would be a matter he could be indifferent to. He believed, as Wang (1974) reports, that the brain functions like a digital computer, but that “*the generally accepted view that there is no mind separated from matter is a false prejudice of our time*”. (p. 326) In this respect his opinion about minds and machines stands in striking contrast with the functionalism advocated by Turing (1950) and Putnam (1960).

2. Lucas

Lucas (1961) has led to a wrong way to apply Gödel's theorem to the problem of minds and machines. His purpose was to defend the view that Gödel's theorem implies the falsity of mechanism as a theory of the mind.

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 is incapable of being produced as 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. (p. 113)

His problem was: Can we justify some kind of anti-mechanism with the help of the incompleteness theorems? And he answered yes to the problem. I will make out in this section the case against this kind of justification of anti-mechanism.

A safeguard against the vulgar inclination in favor of a Lucas-style argument is that pointed out by Hilary Putnam (1960) in his comments on E. Nagel and J.R. Newman (1958). To be able to assert that the mind surpasses any machine, the human mind H has to be capable of discovering a proposition U which no Turing machine M can prove. However, what H can discover is a proposition U such that H can demonstrate:

(*) If M is consistent, then U is undecidable, though true.

It is true that M cannot prove U , but neither can H prove U , unless H can prove that M is consistent. But H cannot prove the consistency of M , if M is very complicated. Moreover, M can perfectly well prove (*). (p. 366)

The following comment by George Boolos (1995) will show that Putnam's criticism is the vital point of the issue even today.

Gödel's disjunctive conclusions concerning the significance of his incompleteness theorems stand in contrast with the conclusion drawn by writers such as E. Nagel and J.R. Newman (1958), J.R. Lucas (1961) and R. Penrose (1989) to the effect that the theorems show outright that the mind is not a Turing machine. The classic reply to these views was given by H. Putnam (1960): Merely to find from a given machine M a statement U for which it can be proved that M , if consistent, cannot prove U is not to prove U —even if M is consistent. It is fair to say that the arguments of these writers have as yet obtained little credence. (pp. 294–5)

Now back to Lucas. His argument was schematically formulated by Rudolf Rucker (1982) as follows :

- L1. If $M^* \subseteq H^*$, then H can see that M embodies a true formal system.
- L2. If H knows that M^* is true, then H knows that $\text{Con}(M) \in H^*$.
- L3. The second incompleteness theorem tells us that $\text{Con}(M) \overline{\in} M^*$.
- L4. If $M^* \subseteq H^*$, then $M^* \neq H^*$.
- L5. If $M^* \supset H^*$, then $M^* \neq H^*$. (pp. 292-3)

Here, H^* is the set of (the Gödel numbers of) sentences which human mathematical intuition can demonstrate, and M^* is the set of (the Gödel numbers of) theorems which a Turing machine yields. " $\text{Con}(\)$ " is supposed to be the number-theoretic predicate which represents "is consistent".

The fallaciousness of Lucas's arguments arises from his presumptive use of human mathematical intuition. Suppose, instead of L1, that

- [1] if $H^* \subseteq M^*$, then M can see that H embodies a true formal system.

This statement has no meaning, because we do not know what is meant by "a Turing machine that can *see* something". This criticism applies to L2 as well if we alternate H with M , because we do not know what "a Turing machine *knows* something" means. L3 has to be corrected as follows :

- [3] The second incompleteness theorem tells us that *if M^* is consistent*, then $\text{Con}(M) \overline{\in} M^*$.

From Lucas's fallacious presumptions it follows that "If $M^* \subseteq H^*$, then $M^* \neq H^*$ ", which excludes the other alternative of G(1), i.e. the possibility that "there may exist a theorem-proving machine which in fact is equivalent to the human mind".

Rucker (1982), however, saw in Lucas's arguments "the *correct* intuition underlying this *fallacious* argument"[Rucker's italics] and modified it as follows.

- R1. [H -Consciousness] $M_e^* \subseteq H^*$ & e is humanly nameable $\Rightarrow \text{Tr}(e) \in H^*$.
 - R2. [H -Platonism] $\text{Tr}(e) \in H^* \Rightarrow \text{Con}(e) \in H^*$.
 - R3. [Gödel's second theorem] [If M_e is consistent, then] $\text{Con}(e) \overline{\in} M_e^*$. (p. 293)
- (The phrase in the square brackets is obliterated by Rucker.)

Here, " $\text{Tr}(\)$ " is an extra-mathematical predicate which is intended to mean that the Turing machine M_e (the suffix e is the largest natural number that H can name) lists a set of (Gödel numbers of) mathematical sentences that H believes himself to be thinking of. H -Consciousness expresses H 's belief that he is not just a formal system, but is rather a mathematician discovering mathematical truths. H -Platonism expresses H 's belief that his Tr -predicate is based on an objectively existing, and therefore consistently described, universe of mathematical objects.

$Tr(e)$ will actually be in H^* , Rucker concludes, only if H can ever be in a position to see M_e as a whole, i.e. only if H is capable of naming the large natural number e .

One alternative which Lucas overlooked was the second part of $G(1)$'s disjunctive conclusion, i.e. the case of $M^* = H^*$. Now let us put $M^* = M_h^*$ if and only if $H^* \subseteq M^*$ holds. The suffix h is the index of M which is equivalent to the mathematical intuition H . If h is humanly unnameable, Rucker says, it does not follow that $Tr(h) \in H^*$, i.e., we cannot know whether $Con(h)$ is true or not. What does it mean that h is humanly unnameable? It means nothing other than that H can never see M_h as a whole. Suppose that there is a machine M_h that is equivalent to human mathematical intuition. Being a theorem-proving machine, M_h is subject to Gödel's second theorem. Therefore, if M_h is consistent, M_h as well as H cannot prove $Con(h)$. This shows that $H^* = M^*$ is compatible with Gödel's second theorem, contrary to Lucas's argument. Thus Rucker (1982) says:

This can only be because the humans can never understand the finite description of M_h well enough to prove that M_h is, in fact, equivalent to their own mathematical intuitions, which are known, by a sort of higher-level intuition, to be consistent. (p. 181)

What is it that Rucker has actually demonstrated? As H -Platonism is weaker than Lucas's premise L2, R2 adds substantially nothing to L2. In fact, the Tr -predicate has no place in Gödel's theorem. " $Tr()$ " cannot in any way be defined in the formal system under consideration. Therefore, the meaning of " $Con(h) \in H^*$ " in H -Platonism is different from that which appears in R3 [Gödel's second theorem]. If " $Con(h) \in H^*$ " means the same in both of its occurrences, it follows that $Con(h) \in H^*$ (from R2) and that $Con(h) \notin H^*$ (from R3), which is a contradiction. Rucker is very ambiguous on the role H -Platonism is to play. On the one hand, it serves to give us the doubtful impression that the human mind surpasses any machine. On the other, its role is to warrant the consistency of M_h . And what he really did by means of H -Platonism was, in essence, just to assert that we *know* an undecidable propositions which a Turing machine cannot yield, which is nothing other than what Lucas had done. H -Platonism does not serve the purpose of substantiating "the *correct* intuition underlying Lucas's *fallacious* argument". Lucas's argument is simply fallacious and involves no *correct* intuition.

Rucker's real modification consists in setting the condition ("if e is humanly nameable") for " $Tr(e) \in H^*$ " to hold. The reason he required H -Consciousness is the same as why he required H -Platonism, the requirement that the antecedent of Gödel's second theorem be satisfied. Without the nameability condition, there would be, contrary to Gödel's opinion, no possibility of the equivalence of the

human mind to a finite machine. We can see now that his conclusion does result not just from Gödel's theorems alone, but from Gödel's second theorem *plus* *H-Consciousness*. Rucker's modification of Lucas's argument is useful only because it made explicit the underlying presumptions of the Lucas-style argument. The way in which he used, or rather misused, the Gödel theorem is fallacious, even if his conclusion appears to be in accord with that of Gödel.

An important origin of Rucker's fallacious argument is, as far as I can see, that he mistook Gödel's philosophical standpoint for a premise of Gödel's reasoning about the implications of the incompleteness theorems. As we have seen above, Gödel's reasoning depends heavily on the contraposition of "objective mathematics" with "subjective mathematics", but not on the differences between the human mind and a Turing machine. Gödel's Gibbs Lecture, titled "Some basic theorems on the foundations of mathematics and their implications", was delivered in 1951, but published only 1995. Until then Wang (1974) had been the only source that could be consulted about Gödel's opinion on this matter. I suspect that Rucker (1984) might have been misled partly by Wang's misrepresentation of Gödel (*1951), specifically by his having adopted the concept "mathematical intuition" where it should be put simply as "the human mind".¹

3. Putnam

Gödel's separation of mind and matter is not the last word, but a problem by itself. However, it suggests something we can learn about the mind, at least in the respect that the mind is not physical-chemical. Functionalism holds that our minds are analogous to computers, and our mental states are simply our "functional states". By "functional states" Putnam originally meant Turing machine states, and the machine table was a normal form for the functional description of the mind. (This is the answer to our Question 2.) In Putnam (1973) he explicitly declares :

What is the importance of machines in the philosophy of mind?Machines forced us to distinguish between an abstract structure and its concrete realization. Not that that distinction came into the world for the first time with machines. But in the case of computing machines, we could not avoid rubbing our noses against the fact that we had to count as to all intents and purposes the same structure could be realized in a bewilderingly variety of different ways ; that the important properties were not physical-chemical. (pp. 299-300)

When Putnam (1960) pointed out the mistake made by Nagel and Newman (1958), his purpose was to defend the thesis that a Turing machine can serve as a model for the human mind. But Putnam (1988), who has taken a turn from

Realism to realism and emerged as a stern critic against functionalism, is not the Putnam of nearly thirty years ago. Whereas Putnam (1960) is a manifesto of functionalism, Putnam (1988) is a detailed criticism of it. It is not the place to review all the reasons that led him to abandon functionalism, but we cannot help noticing that in his arguments against functionalism he makes use of the techniques one uses to prove the Gödel theorem. He says in the last chapter of Putnam (1988) :

A proof in the ordinary sense is a proof in a system which is not just sound, but which a mathematician could, upon reflection, see to be sound, one which a reasonable mathematician would be justified in accepting.

Can this notion of proof itself be formalized ? The question is delicate, but the answer practically all logicians would accept is the following: if there exists a sound system which does formalize this notion of proof, then *that system as a whole* is *not* one that a mathematician could, upon reflection, see to be sound! (p. 117)

What Gödel showed is, so to speak, that we cannot fully formalize our own mathematical capacity *because it is part of that mathematical capacity itself that it can go beyond whatever it can formalize*. Similarly, my extension of Gödelian technique to inductive logics showed that it is part of our notion of justification in general (not just of our notion of *mathematical* justification) that *reason can go beyond whatever reason can formalize*. (p. 118)

In so far as the argument goes, Putnam (1988) looks to prefer the first alternative of G(1), which states that the human mind is incapable of formulating (mechanizing) all its mathematical intuitions, to the second. What has become of his attitude toward the Lucas-style argument ? In his review (1995) of Penrose (1994) he says :

He[Penrose] is convinced by an argument that all experts in mathematical logic have long rejected as fallacious. The fact that the experts all reject Lucas's infamous argument count for nothing in Penrose's eyes. He mistakenly believes that he has a philosophical disagreement with the logical community, when in fact this is a straightforward case of a mathematical fallacy. (p. 370)

It was in order to slide over precisely the possibility that our brains follow a program that is not "simple enough" for us to appreciate in a perfectly conscious way that Lucas blurred the question: exactly what consistency statement was he claiming a human mathematician can prove and a machine cannot ? Penrose completely misses this possibility. (p. 372)

Penrose sketches the proof that any program for generating theorems of mathematics is equivalent to a formal system, but the proof involves describing the entire program in a formal language. Thus if the program is too complicated to appreciate in a perfectly conscious way, the resulting formal system will also be too complicated to understand, and there will be nothing “essentially dubious” about the fact that we cannot prove its soundness. (p. 373)

Putnam retains the second alternative of G(1) as firmly as before. There certainly exists the possibility that the human mind is the product of an algorithmic process which, however, cannot be rendered sufficiently perspicuous for us to know its working. Gödel's theorem has no partiality toward minds, nor toward machines. As a logico-mathematical theorem, the world it applies to is the world of possibilities. In order to eliminate the second alternative, something more is necessary. What we need is some anthropocentric, epistemic principles which would make us abandon the God's-eye view of the human mind—the view that there is a theory of the working of the human mind even *if we cannot recognize it*. Reviewing the reason that led him to abandon functionalism, Putnam (1994) writes :

What did occur to me is this : what the Gödel theorem shows is that human reason is unable to survey itself well enough to see its own limits.

It suggests to me that, first of all, if there is an ‘ideal psychological theory’, that is a theory that does everything that the functionalists want a ‘description of human functional organization’ to do, then there is no reason to believe that it would be within the capacity of human beings to discover it. But, secondly, the *property* of being a ‘description of human functional organization’ is itself such an unclear property that the idea that *there is such a description even if we cannot recognize it* surely goes beyond the bounds of sense. (p. 512) [my italics]

What the second alternative of G(1) means is, in effect, the possibility that there may exist a certain description of “human functional organization” *even if we cannot recognize it*. It is one thing to say that such and such a fact follows from Gödel's theorem, but quite another to use the fact in an argument against functionalism. A logico-mathematical fact cannot decide an issue in the philosophy of mind. In this respect, and only in this respect, Wittgenstein (1978) is justified in saying “a piece of mathematics cannot solve problems of the sort that trouble us”. (p. 388) In the philosophy of mind we have no option but to choose the first alternative of the disjunction of G(1). (Answer to our Question 3).

4. Conclusion³

Princess Margaret defied the royal norm when she fell in love with Group

Captain Peter Taunsend. Eventually, she was forced to choose between Taunsend and the Royal Family. Are we not forced to defy the royal norm of Logic when we claim that certain philosophical consequences follow from Gödel's theorem? In each case study in the preceding chapters on minds and machines, the argument for the alleged philosophical consequences depends on a crucial philosophical premiss. Paul Benacerraf (1998) has called the task of identifying and justifying that premiss the "Princess Margaret problem", and the required premiss the "Princess Margaret premiss". (p. 60) In this essay we have engaged in the game of that problem.

Let G be Gödel's theorem, P the Princess Margaret premiss, and C the alleged philosophical consequences. Then our game consists in how to justify that G and P together entail C , for short: $G, P \vdash C$. A cardinal rule of the game is that P should not be strong enough to entail C , for otherwise the game would be fixed, (i.e., one could prove $P \vdash C$ without making use of G .—Actually P and G are entangled, though.) Take for example, as C , Lucas's conclusion that minds are essentially different from machines, and then the rule requires at least (1) that the human mind should have no special talents that make it surpass all machines, excepting (2) that the human mind has the ability to use language *as part of our natural history*. (Notice that the game is a kind of *solitaire* played by a man or a woman, and *not* a game played between minds and machines. I pretend to see no reason, pace René Descartes, to count (2) as convincing evidence against (1)). A typical misuse is made of Gödel's theorem whenever the argument in question breaches one or more of those rules, which we have ascertained in the Lucas-style argument. I scent a similar misuse in Gödel's argument in favor of $G(2)$.

I believe that all philosophical consequences that do not misuse Gödel's theorem (in my sense of "misuse") have to take the form of disjunction of $G(1)$. The Princess Margaret premiss of $G(1)$ was the "subjective mathematics" in the sense that Gödel defined. To a possible objection that the premiss is too narrow for mathematics to contain the undecidable sentence and that, hence, the "subjective mathematics" is self-contradictory, I want to take notice of the fact that it extends itself indefinitely by taking the undecidable sentence as a new axiom of the extended system. To justify Lucas's preference of human minds over machines, one have to expell the other alternative of $G(1)$, which in turn requires another premiss. The required premiss, however, does not belong to the game in which we have engaged here, but to the philosophy of mind. If there were no Princess Margaret premiss that would not breach any of those rules (although I believe there is), then Gödel's theorem would have nothing to do with the "minds and machines" problem, no matter what love affair may happen to a Margaret.

Notes

- 1 It does not mean that Gödel did not use the term "mathematical intuition" in this context. He actually did when he was criticizing the standpoint of Logical Syntax as follows: "Of course there is no question that by applying mechanically certain rules for handling symbols one arrives at the same sentences as by using mathematical intuition, except that *all* mathematical intuition demonstrably cannot be expressed in *one* finite system of intuitively evident axioms." (Gödel (1953*), p. 340) But, by "mathematical intuition" one usually understands something like a perception, if there were one, of the object of some mathematical theory. This way of using this term stems from his discussion on the continuum hypothesis. The objects of transfinite set theory do not belong to the physical world and even their indirect connection with physical experience is very loose. "But, despite their remoteness from sense experience, we do have something like a perception of the objects of set theory," which according to him is "another kind of relationship between ourselves and reality". (Gödel (1964*), p. 268).
 As for the usage of the term "human mind" Wang (1996) comments as follows: "There is a terminological complication in Gödel's use of the terms *human mind* and *mathematical intuition*. I tend to think in terms of the collective experience of the human species, and so I asked him once about his usage. His reply suggests to me a simplifying idealization. '6.1.23 By mind I mean an individual mind of unlimited life span.'" (p. 189). But, no mention is made here to the term "mathematical intuition."
- 2 If one could reformulate G (2) as follows: "either subjective mathematics surpasses the capability of all computers, or else objective mathematics surpasses subjective mathematics, or both alternatives may be true" (Wang (1996), p. 186), then the second alternative should be a direct consequence of the fact that objective mathematics is not equivalent to subjective mathematics, and that no false proposition can be demonstrated. However, to those who believes that there could be no mathematical truth outside of demonstrability, the distinction of objective and subjective mathematics is itself a Platonist presupposition.
- 3 I am grateful to an anonymous referee for urging me to answer the questions: (1) Has Gödel's philosophical argument anything to do with his incompleteness theorems? (2) What is Gödel's (or my) definition of "the human mind"? This short chapter is an addendum, in which I have tried to explain *my* point. As for Gödel's philosophical standpoint, I disagree, especially on his view of "mind" (so far as I can make sense of it).

References

- Benacerraf, P. (1998) What mathematical truth could not be—I, in M. Schirn ed. *The Philosophy of Mathematics Today*, Clarendon Press, 33–75.
- Boolos, G. (1995) Introductory note to * 1951, in Gödel (1995), 290–304.
- Bowie, G.L. (1982) Lucas' number is finally up, *Journal of Philosophical Logic*, 11, 279–285.
- Davis, M. (1993) How subtle is Gödel's theorem? More on Roger Penrose, *Behavioral and Brain Sciences*, 16, 611–612.
- Gödel, K. (*1951) Some basic theorems on the foundations of mathematics and their implications, in Gödel (1995), 314–323.
- Gödel, K. (*1953/9) Is mathematics syntax of language?, in Gödel (1995), 334–362.
- Gödel, K. (*1964) What is Cantor's Continuum Problem?, in Gödel (1990), 254–270.
- Gödel, K. (1990) *Collected Works, Vol. II: Publications 1938–1974*, Oxford University Press.

- Gödel, K. (1995) *Collected Works, Vol. III : Unpublished Essays and Lectures*, Oxford University Press.
- Lucas, J.R. (1961) Minds, machines and Gödel, *Philosophy*, **36**, 112-127.
- Nagel, E. and Newman, J.R. (1958) *Gödel's Proof*, New York University Press.
- Penrose, R. (1989) *The Emperor's New Mind*, Oxford University Press.
- Penrose, R. (1993) Author's response: An emperor still without mind. *Behavioral and Brain Sciences*, **16**, 616-622.
- Penrose, R. (1994) *Shadows of the mind*, Oxford University Press.
- Putnam, H. (1960) Minds and machines, in Putnam (1975), 362-385.
- Putnam, H. (1973) Philosophy and our mental life, in Putnam (1975), 291-303.
- Putnam, H. (1975) *Mind, Language and Reality : Philosophical Papers, Vol. 2*, Cambridge University Press.
- Putnam, H. (1988) *Representations and Reality*, MIT Press.
- Putnam, H. (1994) Putnam, Hilary (an article written by himself), in S. Guttenplan (ed.), *A companion to the philosophy of mind*, Blackwell, 507-513.
- Putnam, H. (1995) Review of Penrose (1994), *Bulletin of the American Mathematical Society*, **32**, no. 3, 370-373.
- Rucker, R. (1982) *Infinity and the Mind*, Birkhauser.
- Turing, A. (1950) Computing machinery and intelligence, *Mind*, **59**, No. 236, 433-460.
- Wang, H. (1974) *From Mathematics to Philosophy*, Routledge & Kegan Paul.
- Wang, H. (1996) *A Logical Journey : From Gödel to Philosophy*, MIT Press.
- Wittgenstein, L. (1978) *Remarks on the Foundations of Mathematics*, 3rd ed., Basil Blackwell.