E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

University of Economics
Prague

Gödel on the mathematician's mind and Turing Machines

Inês Hipólito
Institute of Philosophy of Nova
Nova University of Lisbon

The choice of a ... machine involves intuition, ... or as [an] alternative one may go straight for the proof and this again requires intuition.

— Letter from Turing to Newman (1940).

Abstract

Gödel's incompleteness theorems are categorically among the most important discoveries ever made not only to Mathematics and logics, but also to Philosophy. Gödel's incompleteness theorems can be applied to demonstrate that the human mind overtakes any mechanism or formal system. Anti-mechanism theses from the incompleteness theorems were presented in Gödel's Proof by Nagel and Newman (1958). Subsequently, J. R. Lucas (1961) claimed that Gödel's incompleteness theorem "proves that mechanism is false, that is, that minds cannot be explained as machines". Furthermore, given any machine which is consistent and capable of doing simple arithmetic, there is a formula it is incapable of producing as being true ...but which we can see to be true" (1961). Moreover, "if the proof of the falsity of mechanism is valid, it is of the greatest consequence for the whole of philosophy" (1961).

More recently, a similar claim has been made by Roger Penrose (1990, 1994) and by Crispin (1994, 1995) in an intuitionist perspective. Generally speaking, all of these support that Gödel's theorems imply, without qualifications, that the human mind infinitely surpasses the power of any finite machine. In the light of this thesis, I would like to consider Gödel's own perspective on an anti-

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

University of Economics
Prague

mechanical thesis. Would Gödel support the thesis that the mathematician mind could be a Turing Machine? What did Gödel think that his theorem could imply about the mathematician's mind? Could the mathematician's mind be a Turing Machine? I will start this discussion with a short review on Lucas and Penrose's arguments, and subsequently I will explore Gödel's own considerations on the disjunction between mathematicians' mind and *Turing machines*.

Introduction

The computational theory of the mind holds that the mind is literally a digital computer. This thesis has been proposed by Hilary Putnam (1960) and developed by Jerry Fodor (1975, 1980, 1987). It combines an account of reasoning with an account of mental states. These latter concerns the representational theory of Mind (RTM) and it supports that intentional states such as beliefs and desires are relations between a thinker and symbolic representations of the content of the state.

The thesis about reasoning — Computational Account of Reasoning (CAR), depends essentially upon this prior claim that intentional states involve symbolic representations. This representations have both semantic and syntactic properties, and processes of reasoning are performed in ways responsive only to the syntax of the symbols — a type of process that meets a technical definition of "computation", and is known as formal symbol manipulation. In this assumption, the mind receives inputs from the perception, maintains and stores this data through memory, handles them via thought and reasoning and generates action by output. Furthermore, the mind is a computational process that is extracted from a hardware/software system in which cognition is the computation executed over mental representations.

In recent years there has been on-going discussion concerning whether Gödel's incompleteness theorems show that the mind is more than a simple machine. This is the anti-mechanism argument that claim that there is at least one thing that the human mind can do that a computer cannot: the human can *see* that the *Gödel Sentence* is true but a machine could not have this *insight*, since the machine must always follow the rules as a formal system. There is a deluge of

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

VŠE
University of Economics
Prague

articles concerning the *mechanist*[1] — *non-mechanist* discussion. Authors of anti-mechanism include J. R. Lucas (1961; 1968; 1970; 1976) and Roger Penrose (1996). I will start by discussing whether the mathematician's mind could be seen as a Turing Machine in Gödel's thought by assessing both Lucas and Penrose's claims, who aimed to proclaim a proof against a mechanical perspective of the human mind through Gödel's incompleteness theorems.

Kurt Gödel's first incompleteness theorem (1931) shows that any consistent formal system in which a "moderate amount of number theory" can be proven will be incomplete, that is, there will be at least one true mathematical claim that cannot be proven within the system (1981, p. 19). The Gödel sentence assets of itself: "I am not provable in *S*," where "S" is the relevant formal system. Suppose that the *Gödel sentence* can be proven in *S*. If so, then by soundness the sentence is true in *S*. But the sentence claims that it is not provable, so it must be that we cannot prove it in *S*. For this reason, the statement that Gödel sentence is provable in S leads to contradiction, so if *S* is consistent, it must be that the Gödel sentence is unprovable in *S*, and therefore true — since the sentence claims precisely that it is not provable. In a nutshell, if consistent, *S* is incomplete.

1. Lucas's Argument: *the human mind is not a Turing Machine*

In his article *Minds, Machines and Gödel* (1961), J. R. Lucas presents a controversial anti-mechanism argument: the argument claims that Gödel's incompleteness theorem shows that the human mind is not a Turing Machine, this is, a computer.

The well known computational theory of the mind is false if Luca's argument succeeds. Furthermore, if Luca's argument is correct, then "strong artificial intelligence" argument, the perspective that it is possible in principle to construct a machine that has the same cognitive abilities as human, is false. In Lucas' words,

> I do not offer a simple knock-down proof that minds are inherently better than machines, but a schema for constructing a *dis*proof of any plausible mechanist thesis that might be proposed. The disproof depends on the particular mechanist thesis

[1] Some authors are Benacerraf (1967); Damjan (1997); Bruni (2006); Chalmes (1996);

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

University of Economics
Prague

> being maintained, and does not claim to show that the mind is uniformly better than the purported mechanist representation of it, but only that it is one respect better and therefore different. That is enough to refute that particular mechanist thesis (Lucas, 1990).

Gödel's proof is at the centre of Lucas's argument. He starts by considering a machine constructed to produce theorems of arithmetic, and argues that regarding operations, this machine would be analogous to a formal system: "if there are only a definite number of types of operation and initial assumptions build into the [machine], we can represent them all by suitable symbols written down on paper" (Lucas, 1961, p.115). On the one hand, we can associate symbols with the specific states of the machine, on the other, one may associate rules of inference with the operations the machine can do to go from one state to the another: "given enough time, paper, and patience, [we could] write down an analogue of the machine's operations", and "this analogue would in fact be a formal proof" (Lucas, 1961, p. 115). This means that the outcome proof that is formalized by a machine will "correspond to the theorems that can be proved in the corresponding formal system" (Lucas, 1961, p.115).

This means that if we would try to prove Gödel's sentence in this formal system, the machine will be unable to produce this sentence as a truth of arithmetic. However, a human mind knows that the sentence is true, and for this reason one may acknowledge that there is at least one thing that the human mind can do that the machine cannot. Therefore, it seems coherent to assent that "a machine cannot be a complete and adequate model of the mind" (Lucas, 1961, p. 113).

If a mechanist formulates a specific mechanistic thesis by claiming, for example, that the human mind is a Turing machine with a given formal specification *S*. Lucas would then refute this thesis by producing *S*'s Gödel sentence, which we can see is true, but the Turing machine cannot. Then, a mechanist can bring forth a different thesis by claiming, for example, that the human mind is a Turing machine with formal specification *S'*. But then Lucas produces the Gödel sentence for *S'*, and so on, until, presumably, the mechanist simply gives up.

If Luca's argument succeeds then the Computational Theory of the Mind and the "strong artificial intelligence" argument are both false, and, therefore, it is impossible to construct a machine that can perfectly emulate our cognitive abilities.

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

VŠE
University of Economics
Prague

Furthermore, if Lucas's argument is true, then the functionalist philosophical perspective of the mind is false.

However, multiple objections came up against Lucas's argument, some of them involving the consistency and inconsistency of the human mind[2]. Some goes as follows: if we cannot establish that human minds are consistent, or if we can establish that they are in fact inconsistent, then Lucas's argument fails. That is, the Gödel sentence will be *true* and *unprovable* only in consistent systems. In an inconsistent system, one can prove any claim whatsoever because in classical logic, any and all claims follow from a contradiction — an inconsistent system will not be complete. But if the machine in question is inconsistent, the machine will be able to prove the Gödel sentence. However, Lucas's argument to succeed, human minds must be consistent. Nevertheless, Gödel's second incompleteness theorem claims that one cannot prove the consistency of a formal system *S* from within the system itself, so, if we are formal systems, we cannot establish our own consistency. For this reason, a mechanist may simple claim that minds are formal systems and therefore, following Gödel's second incompleteness theorem, one cannot establish our own consistency (Hutton, 1976). To sum up, for Lucas's argument to succeed, we must be assured that humans are consistent, while, at the same time, humans cannot ever establish their own consistency.

Other possible objection to Lucas's claim would be to simple deny consistency to the human mind[3]. If humans are inconsistent, then they would be the equivalent to inconsistent Turing Machines, that is, we might be Turing Machines.

In this context, we are left with only two possibilities: (1) humans cannot establish their own consistency, whether they are consistent or not, and (2) humans are in fact inconsistent. However, Lucas cannot support the view that the human mind is inconsistent: if humans were inconsistent machines, humans would potentially endorse any proposition whatsoever. We could also argue that perhaps the inconsistency in question is hidden, buried deep within our belief system; if we are not aware of the inconsistency, then perhaps we cannot use the inconsistency to infer anything at all (Lucas himself mentions this possibility in his (1990)).

---

[3] See Whiteley, 1962.

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

University of Economics
Prague

Other objections involve the problem of the "idealization", (Boyer, 1983; Coder, 1969; Dennett, 1972), since Lucas's scenario involves a hypothetical mind and machine, neither of which being subjected to limitations such as mortality or the inability for some humans to understand Gödel's theorem. Lucas replies that what is really at issue is what can be done by a human and a machine. If, in principle, the human mind can do something that a machine cannot, then the human mind is not a machine, even if it just so happens that any particular human mind could be modeled by a machine as a result of human finitude (Lucas, 1990).

2. Penrose's Argument: *the human mind cannot be computable*

Lucas's argument was revitalized when the physicist R. Penrose formulated and defended a version of it in two books, *The Emperor's New Mind* (1989), and *Shadows of the Mind* (1994). Although there are similarities between Lucas and Penrose's arguments, there are also significant dissimilarities.

Penrose's argument tries to show that the human mind cannot be computable, and this is "the central (new) core argument against the computational modelling of mathematical understanding" (1994).

As reported by Chalmers (1995), Penrose's argument goes as follows:

(1) Suppose that "my reasoning powers are captured by some formal system *F*," and, given this assumption, "consider the class of statements I can know to be true."

(2) Since I know that I am sound, *F* is sound, and so is *F'*, which is simply *F* plus the assumption (made in (1)) that I am *F* (incidentally, a sound formal system is one in which only valid arguments can be proven).

(3) But then "I know that G(*F'*) is true, where this is the Gödel sentence of the system *F'*". However,

(4) Gödel's first incompleteness theorem shows that *F'* could not see that the Gödel sentence is true. Further, we can infer that

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

University of Economics
Prague

(5) I am *F'* (since *F'* is merely *F* plus the assumption made in (1) that I am *F*), and we can also infer that I can see the truth of the Gödel sentence (and therefore given that we are *F'*, *F'* can see the truth of the Gödel sentence). That is,

(6) We have reached a contradiction (*F'* can both see the truth of the Gödel sentence and cannot see the truth of the Gödel sentence).

(7) Therefore, our initial assumption must be false, that is, *F*, or any formal system whatsoever, cannot capture my reasoning powers.

Following Chalmers (1995), there is a great vulnerability with this version in step (2) since the claim that we are sound may lead to contradiction.

Other objections concern the claim (1) and (2). McCullough (1996) claims that for Penrose's argument to succeed, claims (1) and (2) must be true, that is (1) the "Human mathematical reasoning is sound. That is, every statement that a competent human mathematician considers to be 'unassailably true' actually is true", and (2) "the fact that human mathematical reasoning is sound is itself considered to be unassailably true", this claims, however, seem unlikely (McCullough, 1996).

Penrose aims to overcome such objections claiming that there is a distinction between individual, correctable mistakes that mathematicians occasionally make and things they recognize are unassailably true: "If [a] robot is…like a genuine mathematician, although it will still make mistakes from time to time, these mistakes will be correctable…according to its own internal criteria of 'unassailable truth'" (1994). Penrose means that even when mathematicians are fallible, they are still sound because the mistakes are distinguishable from things unassailably true and can be corrected. In other words, mathematicians can make mistakes and still be sound since what matters is the unassailable truth as an output of a sound system.

John Searle (1997), joined the discussion and assuming that a human being can always "see the truth" of a Gödel sentence. More recently McCall (1999), admits that the standard anti-mechanism argument is problematic because the recognition of the truth in Gödel sentence depends essentially on the unproved

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

VŠE
University of Economics
Prague

assumption that the system F under consideration is consistent. McCall's argument claims that human beings — not machines — can see truth and provability.

3. Gödel's Mathematical Intuition

The question that still needs to be addressed is what did Gödel think his first incompleteness theorem implied about mechanism and the mind in general? In his claim, Gödel is much more cautious. In his note to Wang (1972), he explains:

> "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." (Gödel in a note to Wang, 1974[4])

Gödel draws the following inevitable disjunctive conclusion from the incompletes theorems: "either ... 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" (1995). This claim shows that either:

(1) the human mind is not a Turing machine or

(2) there are certain unsolvable mathematical problems.

As reported by Gödel, the second alternative — undecidable mathematical problems — "seems to disprove the view that mathematics is only our own creation; for the creator necessarily knows all properties of his creatures … so this alternative seems to imply that mathematical objects and facts … exist objectively and independently of our mental acts and decisions" (1951). However, Gödel tended to reject the possibility of absolutely unsolvable problems (2). On the other hand, for him to support the first, that the human mind infinitely surpasses any finite machine, would mean to accept the possibility of humanly unsolvable problems.

[4] The note is included in the Collected Works (Gödel, 1990, 305-306). In 1972, Gödel gave Wang a revised version of the note, which Wang published in his From Mathematics to Philosophy (1974, 325-326).

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

University of Economics
Prague

In this assumption, Gödel admits that both mechanism and the alternative that there are absolutely unsolvable problems are consistent with his incompleteness theorems. For philosophical reasons, Gödel struggles in assuming the possibility of the second disjunction, since Gödel thought, inspired by Kant, that the human reason would be fatally irrational if it would ask questions it could not answer. On the other hand, if we are ready to humbly acknowledge human capabilities, and accordingly, admit that there are undecidable mathematical problems, we will naturally reverses Hilbert's optimism; and probably shake Platonism grounds[5].

Lucas has a slightly different perspective (Lucas, 1998), he argues "it is clear that Gödel thought the second disjunct false," that is Gödel "was implicitly denying that any Turing machine could emulate the powers of the human mind." If Lucas is right, it is reasonable to consider that the first thinker to endorse a version of the Lucas-Penrose argument might have been Gödel himself.

The incompleteness results by themselves certainly do not show that the mind is not a computer. The essential extra ingredient that must be added to the incompleteness results is the premise of *rationalistic optimism*: the premise that, as Hilbert famously put it, 'in mathematics there is no *ignorabimus*'—there are no mathematical questions that the human mind is incapable of settling, in principle at any rate, even if this is not so in practice.

> "The incompleteness results do not rule out the possibility that there is a theorem-proving computer which is in fact equivalent to mathematical intuition. ... If my result [incompleteness] 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." (Gödel in conversation with Wang)

The scientific approach to mental phenomena was very important, firstly to Turing and subsequently to Gödel. Gödel's conclusions from his second incompleteness theorem were partly in virtue of Turing's 1936 reduction of finite procedures to machine computations.

Gödel's considerations in Gibbs Lecture, his later conversations with Wang and *Turing's Intelligent Machinery* are an evidence of the attempt to scientifically approach mental phenomena. Both Turing and Gödel were convinced that mental

[5] See Kreisel, 1967.

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

VŠE
University of Economics
Prague

processes were present in mathematical experience. On the one hand Turing noted that for a machine or a brain it is not enough to be converted into a universal (Turing) Machine in order to be intelligent. Therefore, the central scientific task is "to discover the nature of this residue as it occurs in man, and to try and copy it in machines" (Turing, 1948, p. 125).

On the other hand, Gödel considers that there must be a nonmechanical plan to machines, as he reports, "[s]uch a state of affairs would show that there is something nonmechanical in the sense that the overall plan for the historical development of machines is not mechanical. If the general plan is mechanical, then the whole race can be summarised in one machine." (Gödel in *conversation with Wang*).

Conclusions

The incompleteness results by themselves certainly do not show that the mind is not a computer. What needs to be added here is the hilbertian premise that "in mathematics there is no *ignorabimus*", that is, there are no mathematical questions that the human mind is incapable of settling. Penrose seems to turn his back against this premise since, in his perspective, the human mathematical understanding may not be able to solve each problem. Furthermore, he argued that while a formal proof system cannot, because of the theorem, prove its own incompleteness, Gödel-type results are provable by human mathematicians. He takes this disparity to mean that human mathematicians are not describable as formal proof systems, and are not therefore running an algorithm. If the Lucas-Penrose's argument succeeds then the Computational Theory of the Mind and the "strong artificial intelligence" argument are both false, and, therefore, it is impossible to construct a machine that can perfectly emulate our cognitive abilities. Furthermore, if Lucas's argument is true, then the functionalist philosophical perspective of the mind is false. Also, it must be assured that humans are consistent, while they cannot ever establish their own consistency.

Gödel tended to reject the possibility of absolutely unsolvable problems (2). On the other hand, for him to support the first, that the human mind infinitely surpasses any finite machine, would mean that the possibility of humanly unsolvable problems. As he himself explains, "if my result is taken together with

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

University of Economics
Prague

the rationalistic attitude ... 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-theoretical questions undecidable for the human mind".

References

Benacerraf, P. (1967). "God, the Devil, and Gödel," *Monist* 51:9-32.

Boyer, D. (1983). "J. R. Lucas, Kurt Gödel, and Fred Astaire," *Philosophical Quarterly* 33:147-59.

Chalmers, D. J. (1996). "Minds, Machines, and Mathematics," *Psyche* 2:11-20.

Coder, D. (1969). "Gödel's Theorem and Mechanism," *Philosophy* 44:234-7.

Gödel, K. (1995). *Collected Works III* (ed. S. Feferman). New York: Oxford University Press.

Dennett, D.C. and Hofstadter, D. R. (1981). *The Mind's I: Fantasies and Reflections on Self and Soul*. New York: Basic Books.

Hutton, A. (1976). "This Gödel is Killing Me," *Philosophia* 3:135-44.

Lucas, J. R. (1961). "Minds, Machines and Gödel," *Philosophy* 36:112-127.

Lucas, J. R. (1990). "Mind, machines and Gödel: A retrospect." A paper read to the Turing Conference at Brighton on April 6th.

McCall, S. (1999). "Can a Turing Machine Know that the Gödel Sentence is True?" *Journal of Philosophy* 96(10): 525-32.

McCullough, D. (1996). "Can Humans Escape Gödel?" *Psyche* 2:57-65.

Nagel, E. and Newman J.R. (1958). *Gödel's Proof*. New York: New York University Press.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind*. Oxford: Oxford University Press.

E-LOGOS
ELECTRONIC JOURNAL FOR PHILOSOPHY
ISSN 1211-0442 22/2014

University of Economics
Prague

Putnam, H. (1960). “Minds and Machines,” *Dimensions of Mind. A Symposium* (ed. S. Hook). London: Collier-Macmillan.

Searle, J. (1997). Roger Penrose, Kurt Godel, and Cytoskeletons. The Mistery of Consciousness, New York: The New York Review of Books.

Turing, A. (1948). Intelligent Machinery; written in September 1947, submitted to the National Physical Laboratory in 1948; Machine Intelligence *5*, 1969, 3-23.

Wang, H. (1981). *Popular Lectures on Mathematical Logic*. Mineolam NY: Dover.

Wang, Hao. 1974. *From Mathematics to Philosophy*. Humanities Press, New York.

Wright, C. (1994) "About 'The philosophical significance of Godel's theorem': some issues", in Brian McGuinness and Gianluigi Oliver (eds.). *The Philosophy of Michael Dummett,* Kulwer, Dordrecht, 167-202.

Wright, C. (1995) 'Intuitionists are not (Turing) machines', *Philosophia Mathematica* 3,86-102.