Croatian Journal of Philosophy
Vol. XXIV, No. 72, 2024
https://doi.org/10.52685/cjp.24.72.4
Received: November 20, 2023
Accepted: January 15, 2024

Minds, Machines and Gödel

ZVONIMIR ŠIKIĆ
University of Rijeka, Rijeka, Croatia

A very popular argument for the difference between mind and machine are Gödel's incompleteness theorems. Here we present some of the most famous such arguments, as well as their most famous criticisms. Finally, we offer our own reconstruction of the argument and show why it is not valid.

Keywords: Gödel's incompleteness theorems; mind vs. machine; consistency; ω-consistency.

1. *Gödel's theorems*

The vast majority of those who use Gödel's theorems of incompleteness to argue for mind-machine non-equivalence do not fully understand what Gödel's theorems are claiming. So we will begin by presenting the theorems. Gödel's first incompleteness theorem reads as follows.

If formal mathematical theory M includes an appropriate amount of arithmetic it contains an explicitly definable sentence G which asserts its own unprovability and is such that, if M is consistent then $\nvdash_M$ G and if M is ω-consistent then $\nvdash_M$ –G.

In what follows $\vdash$ is $\vdash_M$ and M is a formal mathematical theory which includes an appropriate amount of arithmetic and we think of it as a machine.

Gödel's second incompleteness theorem reads as follows.

If formal theory M is consistent it cannot prove its consistency, Con(M), which is expressed by –Pr ('⊥'), because $\vdash$ Con(M) ↔ G. (About provability predicate Pr (x) see in the appendix.)

Concerning formal unprovability of G and –G, it can be proved that

$\vdash$ –Pr ('G') ↔ Con(M) and $\vdash$ –Pr ('–G') ↔ Con (M + Con(M)).

Notice that Con (M + Con(M)) is stronger than Con(M), by the second incompleteness theorem. On the other hand, it can be proved that Con (M + Con(M)) is a weaker requirement than ω-consistency (even

weaker than 1-consistency which is a weakening of ω-consistency). Ideas of the proofs of some of these results are given in the appendix.

Let us now turn to "Gödelian dualist" arguments and their refutations.

2. *Gödel*

We will start with Gödel. In (Gödel 1951) he admits the possibility that human mind is a machine unable to understand completely its own functioning. By the end of the article I will explain that there are very good reasons for such a point of view.

Gödel even says it is conceivable that it would be known with empirical certainty that the brain suffices for the explanation of all mental phenomena and is a machine in the sense of Turing.

Hence, "Gödelian dualist" would have a hard time convincing Gödel himself.

3. *Penrose, Boolos and Good*

Penrose claims that we can see that G is true as follows (Penrose 1999). If G is provable in Peano arithmetic PA then it is false (because it asserts that it is not provable). But that is impossible *"because our formal system should not be so badly constructed that it actually allows false propositions to be proved* [in other words the system should be correct]*."* So, G is unprovable and therefore true.

Boolos asks what about ZFC (Boolos 1990). If ZFC is correct its Gödel sentence G is also unprovable in ZFC and therefore true. But we don't know if ZFC is correct; "*we could be in the same situation regarding ZFC that Frege was before receiving the letter from Russell."*

Anyway, the argument could be much simpler. If we know that M is correct and therefore consistent then $\vdash \mathrm{Con(M)} \leftrightarrow \mathrm{G}$ implies that we know that G is true. And that's it.

Of course, M also "knows" that, because $\vdash \mathrm{Con(M)} \leftrightarrow \mathrm{G}$.

It could be that we know that Con(M) is true and that we therefore know more than M. But then, we can extend M to M + Con(M) and our knowledge of the truth of Con(M) is then successfully formalized. Of course, now the question is do we know that Con (M + Con (M)) is true etc. The "Gödelian dualist" must verify that the Con sentences of all these extensions are true. But Good successfully argued that no such proof is possible (since it would imply that the smallest non-constructible ordinal is constructible) (Good 1969).

4. *Lucas and Lewis*

Lucas bypasses this hierarchy of extensions (Lucas 1961). Introducing Gödel's theorems, we already said that there is a function Con that assigns a sentence Con(M) to each theory M in such a way that:

C1. Con(M) is true if and only if M is consistent.
C2. If M is correct then Con(M) is true.
C3. Con(M) is provable if and only if M is inconsistent.

Call C a *consistency sentence* for set of sentences S iff there is M such that S is the set of its provable sentences and C = Con(M). Lucas introduced the following rule of inference which is valid, by C2:

L. If C is a consistency sentence for S, infer C from S.

Lucas extended PA to LA, with the rule L. If PA is correct then LA is correct, because L is a valid rule of inference. Furthermore, if LA is a formal theory, its consistency sentence C = Con (LA) would be its theorem, by L, and LA would be inconsistent, by C3. Hence, by C1, the falsehoods would follow from PA. Therefore, if PA is correct we know that Lucas arithmetic is not the output of any formal theory.

So if Lucas can verify all the theorems of Lucas arithmetic then Lucas is no machine.

But we are given no reason to believe that he can. As Lewis warned, in order to check whether Lucas's rule L has been used correctly, a checking procedure would have to decide whether a given set S of sentences is the output of a formal theory and that, we know, is an undecidable problem (Lewis 1989). So we do not know how many theorems of LA Lucas can produce. He can certainly go beyond PA, but he can go beyond it and still be a machine, because limitations on his ability to verify theoremhood in LA may leave him unable to recognize a lot of theorems of LA.

5. *McCall not understanding Godel's theorem*

McCall's reasoning differs from the earlier "Gödelian dualist's" arguments in his admission that the recognition of truth of G, assigned to a formal theory M, depends essentially on the unproved assumption that M is consistent (McCall 1999). That is why McCall refers to the distinction between following formal and informal claims:

A. If M is consistent then G is not provable
   $\vdash$ Con(M) $\rightarrow$ –Pr ('G')
B. If M is consistent then –G is not provable
   $\vdash$ Con(M) $\rightarrow$ –Pr ('–G')).

He claims that both informal sentences are true. He also claims that the formal version of A. is a theorem, whereas the formal version of B. *"to the best of* [*his*] *knowledge"* is not. Hence, McCall concludes that B. yields the informally true but formally unprovable sentence.

But, informal sentence in B. is not true! Unprovability of –G depends on ω-consistency. We can recognize that –G is not provable, if we assume not only the consistency of M, but the ω-consistency of M. And M can do even better, because

$\vdash$ Con (M + Con(M)) $\leftrightarrow$ –Pr ('–G')

and ω-consistency implies Con (M + Con(M)) but is not implied by it (of course, when M proves something then we know it too).

6. *My account*

My own account of dualists' argument is as follows (see Šikić 2005). "Gödelian dualist" argue that no machine M can be identical to a human mathematician H, in the following way. Let M_p be the set of arithmetical sentences provable by M and H_k is to be the set of arithmetical sentences knowable by H (the only property of the notion of knowledge we will need is that knowledge entails truth and that truth does not entail knowledge).

It must be that $M_p \subseteq H_k$ or $M_p \nsubseteq H_k$.

In the second case $M_p \neq H_k$, hence $M \neq H$.

In the first case whatever is provable by M is knowable by H and that means that all sentences in M_p are true. Therefore H knows that M is a correct system. But then H knows that it is a consistent system, i.e. $Con(M) \in H_k$. But $Con(M) \notin M_p$, by second Gödel's theorem, hence $M_p \neq H_k$ and therefore $M \neq H$.

Hence, $M \neq H$ in every case.

But the above conclusion "Therefore H knows that M is a correct system" is not justified. From the truth that every sentence provable by M is knowable by H it follows that every sentence provable by M is true (i.e. that M is correct) but it does not follow that H knows that, because truth does not entail knowledge. It is possible that $M_p \subseteq H_k$ and that H does not know that.

In some specific cases we may know just enough to conclude that M is a correct system. On the other hand, it remains possible that there may exist mathematical machines which in fact are equivalent to our mathematical intuitions. For example, we could be such machines.

What follows from Gödel's incompleteness theorem is that:

There is no machine which could capture all our mathematical intuitions *and which we could understand well enough to know that it is consistent (i.e. that G is true).*

It does not follow that:

There is no machine which could capture all our mathematical intuitions.

We may conclude. As far as Gödel's incompleteness theorems are concerned we could well be machines. But if we are then we are definitely not capable of the complete knowledge of the machines, i.e. of the complete knowledge of ourselves.

That explains Gödel's understanding of the problem in Gödel (1995).

7. *Appendix*

If formal mathematical theory M includes an appropriate amount of arithmetic it can refer to its expression F with its Gödel's number 'F'.

Gödel defined arithmetical predicate Prv (x, y) which represents "x is proved by y" (within M itself) and proved that:

1) n is Gödel's number of a provable formula $\Rightarrow \vdash$ Prv (n, m) for some m
2) n is not Gödel's number of a provable formula $\Rightarrow \vdash$ –Prv (n, m) for every m

Gödel then defined Pr (x), which represents "x is provable", as $\exists y$ Prv (x, y).

Furthermore, (B1) easily follows from 1) and (B1') easily follows from 2) and ω-consistency. It is also easy to prove (B2) and somewhat more difficult (B3).

(B1) $\vdash X \Rightarrow \vdash$ Pr ('X'),
(B1') $\vdash$ Pr ('X') $\Rightarrow \vdash X$ if M is ω-consistent
(B2) $\vdash$ Pr ('X $\rightarrow$ Y') $\rightarrow$ (Pr ('X') $\rightarrow$ Pr ('Y')),
(B3) $\vdash$ Pr ('X') $\rightarrow$ (Pr ('Pr ('X')').

For any predicate P (x), substitution of 'P (d (x))' for x in P (d (x)) gives P (d ('P (d (x))')), or D for short. It immediately follows that D $\Leftrightarrow$ P ('D'). Hence, there is a sentence G such that

(DL) $\vdash G \leftrightarrow$ –Pr ('G')

From (DL), (B1) and (B1') we can deduce the first incompleteness theorem. Namely,

$\vdash G \Rightarrow \vdash$ Pr ('G') $\Leftrightarrow \vdash$–G
$\vdash$ –G $\Leftrightarrow \vdash$ Pr ('G') $\Rightarrow \vdash G$

Both implications contradict the consistency of M. Hence, $\nvdash$G and $\nvdash$–G. Note that we used (B1'), i.e. ω-consistency, to prove the unprovability of –G.

From (DL), (B1), (B2) and (B3) we can deduce the second incompleteness theorem:

$\vdash G \rightarrow$ (Pr ('G') $\rightarrow \perp$)
$\vdash$ Pr ('G') $\rightarrow$ (Pr ('Pr ('G')') $\rightarrow$ Pr ('$\perp$'))
$\vdash$ Pr ('G') $\rightarrow$ Pr ('$\perp$')
$\vdash$ –Pr ('$\perp$') $\rightarrow$ –Pr ('G') i.e. $\vdash$ Con(M) $\rightarrow$ G
$\vdash \perp \rightarrow$ G
$\vdash$ Pr ('$\perp$') $\rightarrow$ Pr ('G')
$\vdash$ –Pr ('G') $\rightarrow$ –Pr ('$\perp$') i.e. $\vdash$ G $\rightarrow$ Con(M)

Now, from $\nvdash$ G and $\vdash$ Con(M) $\leftrightarrow$ G it immediately follows that $\nvdash$ Con(M).

So, by (DL) and $\vdash$ Con(M) $\leftrightarrow$ G, unprovability of G is provably equivalent to the consistency of M:

$\vdash$ –Pr ('G') $\leftrightarrow$ Con(M)

What do we know about the unprovability of –G, which is the other part of the first incompleteness theorem? From $\vdash$ –G $\leftrightarrow$ Pr (), by (B1) and (B2), we get

$\vdash$ –Pr ('–G') $\leftrightarrow$ –Pr ('Pr ('$\perp$')').

But –Pr ('Pr ('⊥')') expresses the consistency of M + Con(M). Namely, if $Pr_{M+Con(M)}$ is the provability predicate of M + Con(M), then the consistency of M + Con(M) is expressed by $–Pr_{M+Con(M)}$ ('⊥'). But,

$$–Pr\ (\text{'Pr}\ (\text{'}\bot\text{'})\text{'}) \Leftrightarrow –Pr_{M}\ (\text{'}–Con(M)\text{'}) \Leftrightarrow –Pr_{M+Con(M)}\ (\text{'}\bot\text{'})$$

Hence

$$\vdash –Pr\ (\text{'}–G\text{'}) \leftrightarrow Con\ (M + Con(M)).$$

References

Boolos, G. 1990. "On seeing the truth of the Gödel sentence." *Behavioural and Brain Sciences* 13: 655–656.

Gödel, K. 1931. "Über formal unentscheidbäre Sätze I." *Monatshefte für Mathematik und Physik* 38: 173–198.

Gödel, K. 1995. "Gibbs Lecture, 1951." In S. Feferman (ed.). *Collected Works, Vol. 3: Unpublished Essays and Lectures*. Oxford: Oxford University Press, 290–323.

Good, I. J. 1969. "Gödel's theorem is a red herring." *The British Journal for the Philosophy of Science* 18: 359–373.

Lewis, D. 1989. "Lucas against mechanism II." *Canadian Journal of Philosophy* 9: 373–376.

Lucas, J. R. 1961. "Minds, machines and Gödel." *Philosophy* 36: 112–137.

McCall, S. 1999. "Can a Turing machine know that the Gödel sentence is true?" *Journal of Philosophy* 96: 525–532.

Penrose, R. 1999. *The Emperor's New Mind*. Oxford: Oxford University Press.

Penrose, R. 1994. *Shadows of the Mind*: *A Search for the Missing Science of Consciousness*. Oxford: Oxford University Press.

Šikić, Z. 2005. "Gödel's Incompleteness and Man-Machine Non-Equivalence." *Grazer Mathematische Berichte* 304: 75–78.