

# Gödel's incompleteness theorem and man-machine non-equivalence

Zvonimir Šikić

## Abstract

We analyse the argument that Gödel's incompleteness theorem implies man-machine non-equivalence, and prove that it is incorrect. Instead, we offer a correct implication.

For many, it is still hard to conceive how the world of subjective experiences spring out of merely physical events. This problem of qualia is the hardest and the main part of the mind-body problem. The problem is often summed up in the following question: "How matter (i.e. body and brain) becomes mind?" All sorts of dualists think it never does and some of them, like Lucas [2] and Penrose [3], think that Gödel's incompleteness theorem proves that. Their main argument is that Gödel's theorem implies man-machine non-equivalence in the following sense:

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

Hence, we are not just machines, there is something beyond that.

We'll start with the bare bones of Gödel's incompleteness. Let  $M$  be a machine which is programmed to print finite sequences of three symbols:  $-$ ,  $P$ ,  $D$ . (In what follows these sequences will be simply called sequences.) At each stage one sequence is printed into a square and each square is part of the tape unending in one direction.

We say that  $M$  prints a sequence if  $M$  prints it at some stage. We say that  $M$  does not print a sequence if  $M$  does not print it at any stage. Some of the sequences are meaningful and we call them sentences. Here is the definition.

**Definition of sentences and their meanings:**

A sentence is a sequence of the form  $PX$ ,  $-PX$ ,  $PDY$  or  $-PDY$ , where  $X$  is any sequence not starting with  $D$  and  $Y$  is any sequence.

1. The meaning of a sentence of the form  $PX$  is "M prints X".
2. The meaning of a sentence of the form  $-PX$  is "M does not print X".
3. The meaning of a sentence of the form  $PDY$  is "M prints YY".
4. The meaning of a sentence of the form  $-PDY$  is "M does not print YY".

**Remark:** It helps to think of  $-$ ,  $P$  and  $D$  as words meaning "not", "prints" and "double".

Sentences have meanings, so they are true or false. All other sequences are neither true nor false, they are meaningless. Our main interest concerning machines are their correctness and completeness. These notions are now easily defined.

**Definition of correctness:**

*Machine M is correct if it prints only true sentences (i.e. if M prints a sentence then this sentence is true).*

**Remark:** A correct machine may print a lot of non-sentences. For example, a machine which prints only non-sentences is trivially correct.

**Definition of completeness:**

*Machine M is complete if it prints all true sentences.*

**Remark:** A complete machine may print a lot of false sentences. For example, a machine which prints all sequences is trivially complete.

The most interesting machines would be those which are correct and complete. Unfortunately, there is no such machine (i.e. it is impossible to construct such a machine).

**Theorem on correctness and completeness impossibility:**

*Every machine is either incorrect or incomplete.*

*Proof.*

Take a look at sentence  $-PD - PD$ . It means "M does not print  $-PD - PD$ ". Of course,  $M$  either prints  $-PD - PD$  or not.

If  $M$  prints  $-PD - PD$  then it prints a false sentence i.e.  $M$  is not correct.

If  $M$  does not print  $-PD - PD$  then it does not print a true sentence i.e.  $M$  is incomplete. □

Now, what is the link between this very simple theorem, Gödel's incompleteness theorem and mechanisation of our mathematical intuitions? To mechanize our mathematical intuitions means to construct a machine which would prove (and then print!) all mathematical theorems which are normally derived using these intuitions. A formalized

mathematical theory with explicitly defined language, axioms and deductive rules is such a machine. Hence, here is the machine to which we may apply our simple theorem. But there is one serious problem. Mathematical machines are not self reflective in the sense that the machines from our theorem are. These machines produce sentences which assert something about the machines themselves. Our mathematical theories/machines assert many things about various mathematical objects, but nothing about the theories/machines themselves. If we are interested in a theory/machine itself we usually construct another theory/machine, called meta-theory/meta-machine, to deal with it. But here comes Gödel. In his famous [1] he proved that a mathematical theory/machine, which includes an appropriate amount of arithmetic, may represent its own meta-theory/meta-machine so that our simple theorem applies. (This is the hardest part of Gödel's proof.) Hence, we have:

**If mechanized mathematical theory includes an appropriate amount of arithmetic it is either incorrect or incomplete, i.e. if it is correct then it is incomplete. Even more, we can explicitly define Gödel's sentence (corresponding to our  $-PD - PD$ ) which is true but not provable in the theory.**

Now, the dualists argument is as follows. Any attempt to mechanize our mathematical intuitions is doomed to fail because the very fact of mechanization yields new intuitive knowledge, e.g. Gödel's sentence, which is not captured by the mechanization.

What is wrong with this argument? The problem is that you have to know quite a lot about the specific mechanization to conclude that its corresponding Gödel's sentence is true. In the specific case that Gödel analysed and that we partially presented above we know just enough to conclude that. On the other hand, it remains possible that there may exist (and even be empirically discoverable) a mathematical machine which in fact is equivalent to our mathematical intuitions. For example, we could be such machines.

So, dualists like Lucas and Penrose confused the incorrect argument:

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

with the correct argument:

**There is no machine which could capture all our mathematical intuitions and which we could understand well enough to see that its Gödel's sentence is true.**

We may conclude. As far as Gödel's incompleteness theorem is 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.

## References

- [1] Gödel, K., *Über formal unentscheidbare Sätze I*, Monatshefte für Mathematik und Physik. *38*, 173–198, 1931.
- [2] Lucas, J.R., *Minds, machines and Gödel*, Philosophy. *36*, 112–137, 1961.
- [3] Penrose, R., *The Emperor's New Mind*, Oxford University Press. 1989.

Zvonimir Šikić  
*Katedra za matematiku i nacrtnu geometriju,*  
*Fakultet strojarstva i brodogradnje,*  
*Ivana Lučića,*  
*Zagreb, Croatia*  
e-mail: [zsikic@fsb.hr](mailto:zsikic@fsb.hr)