STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 81–82
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.04

Note

RUDY RUCKER *

A NOTE ON THE LUCAS ARGUMENT

This note is derived from my books *Infinity and the Mind* (2005, Preface) and *The Lifebox, the Seashell, and the Soul* (2016, footnote 102).

We're talking about J. Anthony Lucas's classic argument that Gödel's Second Incompleteness Theorem rules out man-machine equivalence. This is an argument that Penrose revived and popularized in the 1990s. This fallacious argument is a thoroughly dead horse. But I'll give it another beating here. Do note that the Lucas-Penrose argument is a completely distinct issue from Penrose-Hameroff speculation that the brain can act as a coherent quantum computer. It's to Penrose's credit that he's associated with multiple controversial ideas!

Before continuing, I should explain the Gödel's Second Incompleteness Theorem is the result that if F is a consistent formal system as strong as arithmetic, then F cannot prove the sentence $Con(F)$. $Con(F)$ is the sentence that expresses the consistency of F by asserting that F will never prove, say, $0 = 1$. If we think of h as being the index of the Turing machine *Mh,* we can write $Con(h)$ as shorthand for $Con(Mh)$.

Suppose h is an integer that codes the program for a device *Mh* whose output is very much like a person's. Lucas and Penrose want to say the following

(1) After hanging around with *Mh* for a while, any reasonable person will feel like asserting $Tr(h)$, a sentence which says something like, "If I base a machine *Mh* on the algorithm coded by h I'll get a machine which only ouputs true sentences about mathematics".

(2) Having perceived the truth of $Tr(h)$, any reasonable person will also feel like asserting $Con(h)$, a sentence which says something like, "If I base

* San Jose State University, Department of Computer Science. E-mail: rudy@rudyrucker.com. ORCID: 0000-0001-7679-3025.

a machine *Mh* on the algorithm coded by *h* I'll get a machine which never generates any mathematical contradictions".

(3) Gödel's Second Incompleteness Theorem shows that *Mh* can't prove *Con*(*h*), so now it looks as if any reasonable person who hangs around with a human-like *Mh* will soon know something that the machine itself can't prove.

The philosopher Hilary Putnam formulated what remains the best counterargument in his 1960 essay, *Minds and Machines* (1964). For Lucas's ripostes to such objections, see his genial if unconvincing essay, *A Paper Read to the Turing Conference at Brighton on April 6th, 1990* (Lucas, 1990).

Putnam's point is simple. Even if you have seen *Mh* behaving sensibly for a period of time, you still don't have any firm basis for asserting either that *Mh* will always say only true things about mathematics or that *Mh* will never fall into an inconsistency. Now if you were to have a full understanding of how *Mh* operates, then perhaps you could prove that *Mh* is consistent. But, in the case where *h* is the mind recipe, the operation of the eventual *Mh* is incomprehensibly intricate, and we will never be in a position to legitimately claim to know the truth of the sentence *Con*(*h*) which asserts that *Mh* is consistent. This is, indeed, the content of Gödel's Second Incompleteness Theorem. Rather than ruling out man-machine equivalence, the theorem places limits on what we can know about machines equivalent to ourselves.

And, really, this shouldn't come as a surprise. You can share an office or a house with a person *P* for fifteen years, growing confident in the belief that *P* is consistent, and then one day, *P* begins saying and doing things that are completely insane. You imagined that you knew *Con(P)* to be true, but this was never the case at all. The only solid reason for asserting *Con*(*P*) would have been a systematic proof, but, given that you and *P* were of equivalent sophistication, this kind of proof remained always beyond your powers. All along, the very fact that *Con*(*P*) wasn't provable contained the possibility that it wasn't true. Like it or not, that's the zone we operate in when relating to other intelligent beings.

REFERENCES

Lucas, J. R. (1990). A Paper Read to the Turing Conference at Brighton on April 6th, 1990. Retrieved from: http://users.ox.ac.uk/~jrlucas/Godel/brighton.html

Putnam, H. (1964). Minds and Machines. In: R. Anderson, *Minds and Machines* (pp. 43–59). Upper Saddle River: Prentice-Hall.

Rucker, R. (2005). *Infinity and the Mind* (3rd ed.). Princeton: Princeton University Press.

Rucker, R. (2016). The Lifebox, the Seashell, and the Soul. Edinburgh: Transreal Fiction.