

MECHANISM, TRUTH, AND PENROSE'S NEW ARGUMENT

Received 4 September 2001; received in revised version 22 July 2002

**ABSTRACT.** Sections 3.16 and 3.23 of Roger Penrose's *Shadows of the mind* (Oxford, Oxford University Press, 1994) contain a subtle and intriguing new argument against mechanism, the thesis that the human mind can be accurately modeled by a Turing machine. The argument, based on the incompleteness theorem, is designed to meet standard objections to the original Lucas–Penrose formulations. The new argument, however, seems to invoke an unrestricted truth predicate (and an unrestricted knowability predicate). If so, its premises are inconsistent. The usual ways of restricting the predicates either invalidate Penrose's reasoning or require presuppositions that the mechanist can reject.

**KEY WORDS:** incompleteness, Lucas, mechanism, Penrose, truth

Starting with J. R. Lucas (1961), a much discussed line of argument goes from the incompleteness of arithmetic to the repudiation of the mechanistic thesis that the human mind is, or can be accurately modeled as, a digital computer or a Turing machine. Suppose that a mechanist claims that the output of a particular machine  $M$  consists of all and only the arithmetic truths that a given human, such as Lucas, or any group of humans, will ever or can ever know. Assume that the output of  $M$  consists of only arithmetic truths. Since Lucas understands the proof of Gödel's incompleteness theorem, he can study  $M$  and produce its Gödel sentence  $G$ . Lucas knows that  $G$  will never be produced or "asserted" by  $M$ . He also knows that  $G$  "says" that  $G$  will never be produced by  $M$ . Thus, Lucas knows that  $G$  is true. So the mechanist was mistaken in the claim that the output of  $M$  contains all the truths that (any group containing) Lucas can know. The idea is that the incompleteness theorem provides the resources to refute any particular claim made by a mechanist.

Gödel's correspondence and other writings contain a carefully qualified version of this argument, and the eminent mathematician and physicist Roger Penrose has recently joined in (1989, especially Chapters 4, 10). So the Gödelian anti-mechanists are a powerful intellectual group to reckon with. Over the years, a number of authors have attacked the Lucas–Penrose position. In (1996) Lucas presented an extensive reply to his critics, and about 200 pages of Penrose (1994) are devoted to even more extensive responses to various criticisms of the anti-mechanist argument. The subject of this paper is an intriguing new version of the argument found in Penrose (1994, §§3.16, 3.23) (see also Penrose (1996)).<sup>1</sup>



## 1. WHAT EXACTLY IS THE ISSUE HERE?

As it stands, the bare statement that the human mind can be modeled as a digital computer or Turing machine is too obscure to be adjudicated by a theorem of mathematics. Writing on Gödel's version of the argument, George Boolos (1995, p. 293) pointed out something:

... it is certainly not obvious what it means to say that the human mind, or even the mind of someone human being, *is* a finite machine, e.g., a Turing machine. And to say that the mind (at least in its theorem-proving aspect), or *a* mind, may be represented by a Turing machine is to leave entirely open just *how* it is so represented.

Similarly, Per Lindström (2001) speculated that “it may turn out that [our] question is not well-defined and so has no well-defined answer”.

The mechanist claims that there can be a machine whose outputs are the same as those of a human or a group of humans. The anti-mechanist opponent disputes this, claiming the opposite. What sort of machine are they talking about? What outputs of what aspect of what humans? To make the connection with incompleteness, for “output” we stick to propositions that can be rendered in the language of first-order Peano arithmetic. Penrose (1996) suggests that for the sake of the argument, we can further restrict the focus to  $\Pi_1$ -sentences. The totality of arithmetic or  $\Pi_1$ -sentences that a given person asserts in his lifetime is finite, and this totality is almost certainly inconsistent. All it takes is for the person to make a mistake in calculation. If the mechanist claims that there could be a machine whose output is one of these finite sets, or the truths among one of these sets, or the logical consequences thereof, then the incompleteness theorems are irrelevant.

To get an interesting thesis, and provide an interface for incompleteness, we must idealize on both the computers and the people. The idealizations on the machine side are familiar, similar to idealizations made throughout mathematics. Actual computers have fixed limits on memory and are subject to hardware malfunctions and software bugs. Here we ignore finite limits and assume that our machines never run out of memory or working space. We also assume that they run indefinitely without crashing, always faithfully following their instructions. We make the usual distinction between hardware and software, and then ignore the hardware. In short, we deal with something in the neighborhood of Turing machines.

So far, we have *widened* the gap between human and machine. Our question now concerns the relationship between the pristine realm of Turing machines and the flesh and blood humans we know and love. Rather than deal with the assertions of an actual mathematician, the mechanist and anti-mechanist alike refer to what an ideal human, or the community