

Simulation and Computability: Why Penrose fails to prove the impossibility of Artificial Intelligence (and why we should care)

Ron Chrisley

Director, Centre for Cognitive Science (COGS) c/o
Department of Informatics University of Sussex United Kingdom

ronc@sussex.ac.uk

Abstract

In two widely-read books, Roger Penrose resuscitates an argument (originally considered by Gödel, Turing and Lucas) against the possibility of formalising human thought, and by extension, against the possibility of artificial intelligence (AI). The argument is essentially logico-mathematical, drawing on principles used in Cantor's diagonal argument and Gödel's incompleteness theorem, and concludes that there is a function F (a variation on the halting function) that humans can compute which no Turing machine can. After briefly reviewing the argument, I argue that it fails to support Penrose's claims. Exploiting an insight arrived at independently by Whitley, I contend that while Penrose's formal argument (for the claim that humans can compute F , but no Turing machines can) is sound and valid, the informal, sceptical conclusion concerning AI that he draws from it is a non-sequitur. Specifically, the fact that no Turing machine can compute F , yet humans can, does not imply that there are aspects of human cognition that are not simulable by any Turing machine. I show this by considering an analog of the halting function, defined in terms of humans rather than Turing machines. Seeing how this "person halting problem" refutes Penrose requires re-examining our notions of simulation, computability and function individuation, making the position of relevance to those not concerned with the possibility of artificial intelligence. In particular, the position implies that according to orthodox criteria for individuating functions, there can be no Universal Turing machine (i.e., there can be no Turing machine that can compute all Turing-computable functions). I suggest an alternative means of function individuation that preserves the possibility of Universality, and explore some of its consequences for computability theory.