Review of *Shadows of the Mind* by Roger Penrose

John McCarthy, Stanford University

1 Introduction

This book and its predecessor *The Emperor's New Mind* argue that natural minds cannot be understood and artificial minds cannot be constructed without new physics, about which the book gives some ideas. We have no objection to new physics but don't see it as necessary for artificial intelligence. We see artificial intelligence research as making definite progress on difficult scientific problems. I take it that students of natural intelligence also see present physics as adequate for understanding mind.

This review concerns only some problems with the first part of the book. [1]

2 Awareness and Understanding

Penrose discusses *awareness* and *understanding* briefly and concludes (with no references to the AI literature) that AI researchers have no idea of how to make computer programs with these qualities.

I substantially agree with his characterizations of *awareness* and *understanding* and agree that definitions are not appropriate at the present level of understanding of these phenomena. We disagree about whether computers can have *awareness* and *understanding*.

Here's how it can be done within the framework of *pure logical AI*.

Pure logical AI represents all the program's knowledge and belief by sentences in a language of mathematical logic. Purity is inefficient but makes the discussion brief. ([McCarthy, 1989]) is a general discussion of logical AI and has additional references.

We distinguish a part of the robot's memory, which we will call its *consciousness*. Sentences have to come into consciousness before they are used in reasoning.

Reasoning involves logical deduction and also some *nonmonotonic* reasoning processes. The results of the reasoning re-enter consciousness. Some old sentences in consciousness get crowded out into the main memory.

Deliberate action in a pure logical robot is a consequence of the robot inferring that it should do the action. The actions include external motor and sensory actions (observations) but also *mental actions* such as retrieval of sentences from the general memory into consciousness.

Awareness of the program's environment is accomplished by the automatic appearance of certain class of sentences about the program's environment in the program's *consciousness*.

[1] Considerations in my review [McCarthy, 1990a] of the earlier book are not repeated here.

These sentences often appear through *actions* of *observation* but should often result from built-in observations, e.g. noticing who comes into the room.

Besides awareness of the environment, there is also *self-awareness*. Self-awareness is caused by *events* and *actions* of self-observation including observations of consciousness and of the memory as a whole. The sentences expressing self-awareness also go into consciousness.

The key question about awareness in the design of logical robots concerns what kinds of sentences can and should appear in consciousness—either automatically or as the result of mental actions. Here are some examples of required mental actions.

- Observing its physical body, recognizing the positions of its effectors, noticing the relation of its body to the environment and noticing the values of important internal variables, e.g. the state of its power supply and of its communication channels.

- Observing whether it knows the telephone number of a certain person. Observing that it does know the number or that it can get it by some procedure is likely to be straightforward logical deduction. Inferring that it doesn't know the number and can't get it by reasoning requires getting around Gödel's theorem, because inferring that any sentence does not follow carries with it an implication that the theory is consistent, and Gödel tells us that this cannot be done entirely within a theory.

  Our approach uses Gödel's [Gödel, 1940] notion of relative consistency which allows inferring that if the theory is consistent, then a certain sentence doesn't follow. In cases of main AI interest, this can be done without the complications that Gödel had to introduce in order to prove the consistency of the continuum hypothesis. See ([McCarthy, 1995]) for a start on details.

- Keeping a journal of physical and intellectual events so it can refer to its past beliefs, observations and actions.

- Observing its goal structure and forming sentences about it.

- Observing its own intentions. The robot may *intend* to perform a certain action. This would let it later infer that certain possibilities are irrelevant in view of its intentions.

- Observing how it arrived at its current beliefs. Most of the important beliefs of the system will have been obtained by nonmonotonic reasoning, and are therefore uncertain. It will need to maintain a critical view of these beliefs, i.e. believe meta-sentences about them that will aid in revising them when new information warrants doing so.

- Not only pedigrees of beliefs but other auxiliary information should either be represented as sentences or be observable in such a way as to give rise to sentences. Thus a system should be able to answer the question: "Why don't I believe p?".

- Regarding its entire mental state up to the present as an object, i.e. a context. [McCarthy, 1993] discusses contexts as formal objects. The ability to *transcend* one's present context and think about it as an object is an important form of introspection, especially when we compare human and machine intelligence.

- Knowing what goals it can currently achieve and what its choices are for action. Understanding and reasoning about one's own choices constitutes *free will.*

It seems to me that the notions of awareness and understanding outlined above agree with Penrose's characterizations on p._ 37. However, his ideas about free will strike me as quite confused and not repairable. [McCarthy and Hayes, 1969] discusses free will in deterministic systems, e.g. interacting finite automata.

3 The Argument from Gödel's Theorem

The argument about whether humans necessarily have superior minds to robots is unique among philosophical arguments in getting far into mathematical logical technicalities. This is not Penrose's fault. What machines can and cannot do in principle really is a technical logical question. Here's how it gets messy.

A. Whatever formal axiomatization of arithmetic the robot uses, Gödel's theorem shows how to construct from that axiomatization a sentence that is true if that axiomatization is sound but which cannot be proved in the axiomatization. This can be done in Turing's (1940) way or in Feferman's (1962) way. Both are discussed in [Feferman, 1988].

B. Yes, but the construction of this sentence is accomplished by a program the robot can also apply either to its previous system to get a new one or to a system used by its interlocutor.

A. This process can be iterated through transfinite ordinals, and the ordinals the robot can use will have an upper bound. The human can in principle determine this bound by inspecting the robot's program.

B. To iterate through ordinals requires *ordinal notations.* These are notations for computable predicates, but it is necessary to establish that the computation really produces a well-founded total ordering. Thus we need to consider *provably recursive ordinals.* Then we need to ask what axiomatic system is to be used for these proofs.

Moreover, the new axiomatic systems obtained by the iteration depend on the notation and not merely on the ordinal number the notation determines.

To me, and maybe to Penrose, it is implausible that the possibilities of human thought, except in recursive function theory, can depend strongly on these advanced considerations.

4 Modes of Reasoning

Part of Penrose's conviction that his reasoning is intrinsically more powerful than that of a computer program may come from his using kinds of reasoning that he implicitly denies machines. There are two such kinds of reasoning.

The first is that he reasons about theories in general, i.e. he uses variables ranging over theories. As far as I can see he never allows for the computer program doing that. However, reasoning about theories as objects is not different in principle from reasoning about other objects.

The second is that much of Penrose's reasoning is nonmonotonic, e.g. preferring the simplest explanation of some phenomenon, but his methodology doesn't allow for nonmonotonic reasoning by the program. Mathematicians' acceptance of the axiom of choice, for example, occurs through informal nonmonotonic reasoning. Formalized nonmonotonic reasoning is a recent development.

References

[Abelson and Sussman, 1985] **Abelson, Harold and Gerald Sussman**: *Structure and Interpretation of Computer Programs*, M.I.T. Press, 1985.

[Feferman, 1988] **Feferman, Solomon**: "Turing in the land of $O(z)$. In *The Universal Turing Machine: a Half-century Survey* (ed. R. Herken). Oxford University Press, 1988.

[Gödel, 1940] **Gödel, Kurt**: *The Consistency of The Axiom of Choice and of the Generalized Continuum-Hypothesis with the Axioms of Set Theory.* Princeton University Press, 1940.

[McCarthy and Hayes, 1969] **McCarthy, John and P.J. Hayes**: "Some Philosophical Problems from the Standpoint of Artificial Intelligence", in D. Michie (ed), *Machine Intelligence 4*, American Elsevier, New York, NY, 1969. Reprinted in [McCarthy, 1990].

[McCarthy, 1989] **McCarthy, John (1989)**: "Artificial Intelligence and Logic" in Thomason, Richmond (ed.) *Philosophical Logic and Artificial Intelligence* (Dordrecht ; Kluwer Academic, c1989). Also accessible from http://www-formal.stanford.edu/jmc/home.html.

[McCarthy, 1990] **McCarthy, John (1990)**: *Formalizing Common Sense*, Ablex, Norwood, New Jersey, 1990.

[McCarthy, 1990a] **McCarthy, John (1990a)**: Review of *The Emperor's New Mind* by Roger Penrose, in *Bulletin of the American Mathematical Society*, Volume 23, Number 2, October 1990, pp. 606-616. Also accessible from http://www-formal.stanford.edu/jmc/home.html.

[McCarthy, 1993] **McCarthy, John (1993)**: "Notes on Formalizing Context" IJCAI-93. Morgan-Kauffman. Also accessible from http://www-formal.stanford.edu/jmc/home.html.

[McCarthy, 1995] **McCarthy, John (1995)**: "Making Robots Conscious of their Mental States". Invited lecture at the Symposium on Consciousness, AAAI, Spring 1995. Also accessible from http://www-formal.stanford.edu/jmc/home.html.

/@sail.stanford.edu:/u/jmc/w95/penrose2.tex: begun 1995 Jan 16, latexed 1995 Feb 25 at 2:28 p.m.