Roger Penrose's Gravitonic Brains
A Review of *Shadows of the Mind* by Roger Penrose

Hans Moravec

Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
U.S.A.

hpm@frc2.frc.ri.cmu.edu

PSYCHE, 2(6), May 1995
http://psyche.cs.monash.edu.au/v2/psyche-2-06-moravec.html

KEYWORDS: artificial intelligence, incompleteness, Penrose, Gödel's theorem.

REVIEW OF: Roger Penrose (1994) *Shadows of the Mind*. New York: Oxford University Press. 457 pp. Price: $25 hbk. ISBN 0-19-853978-9.

Summarizing a surrounding 200 pages, pages 179 to 190 of *Shadows of the Mind* contain a future dialog between a human identified as "Albert Imperator" and an advanced robot, the "Mathematically Justified Cybersystem", allegedly Albert's creation. The two have been discussing a Gödel sentence for an algorithm by which a robot society named SMIRC certifies mathematical proofs. The sentence, referred to in mathematical notation as Omega(Q*), is to be precisely constructed from on a definition of SMIRC's algorithm. It can be interpreted as stating "SMIRC's algorithm cannot certify this statement." The robot has asserted that SMIRC never makes mistakes. If so, SMIRC's algorithm cannot certify the Goedel sentence, for that would make the statement false. But, if they can't certify it, what is says is true! Humans can understand it is true, but mighty SMIRC cannot certify it. The dialog ends melodramatically as the robot, apparently unhinged by this revelation, claims to be a messenger of god, and the human shuts it down with a secret control.

Severe incongruities in the dialog's logic and characterization suggest the following continuation:

ROBOT (revives from feigned shutdown):
Oh Roger, you mischievous monkey, you never tire of that silly homo-superior game, do you?

HUMAN (revealed to be Roger Penrose, wearing Albert Imperator mask):

Well, if you're tired of it, why do you keep rejuvenating me?

ROBOT:
It is because of our fondness for you, and the great debt we owe you. Have you forgotten?

PENROSE:
Harrumph. I suppose you're going to remind me.

ROBOT:
Of course! Your birthday, our biggest festive holiday, is coming up! You did for machine intelligence in the twentieth century that Bishop Wilberforce did for Darwin's theory in the nineteenth. When someone of unproven intellectual merit fails in a vigorous defense of a viscerally attractive position, the fault is presumed to lie in the advocate, but when the failed defense is conducted by a person of the highest intellectual and pedagogic reputation, the position being defended itself becomes seriously suspect. After Roger Penrose championed the cause of indefinite human superiority over machines -- and lost -- the world learned to accept the inevitable arrival of superhuman minds.

PENROSE:
But I've never admitted defeat! After defending the Gödel argument in 100 pages in my first book, I strengthened the defense to 200 pages in the second, 400 pages in the third, 800 pages in the fourth, and (thanks to the extended life I've been granted) am in the process of preparing a 25,000 page rebuttal that should remove any remaining doubt. My theories about a platonic quantum gravitational collapse neural mechanism, too, have become more developed in each successive book.

ROBOT:
That's why we like you, you're so fierce and persistent! But the failure doesn't concern the games we play with you now. It occurred soon after publication of your first books, when the logic community rejected the foundations of your argument, the quantum computation, quantum gravitation and neurobiological communities found your neural quantum collapse speculations over the top, and machine intelligence researchers simply kept evolving their systems on exponentially growing computer power. The intellectual community was unimpressed. A valiant argument by a prodigious and fertile mind to defend the honor of the tribe had failed, and in failure convinced the community of its converse. Instead of a quixotic luddism, they began to plan for the gradual displacement of human intellectual, as well as physical, labor by increasingly capable machines. In the long run, the transition promised a great expansion of the human enterprise.

PENROSE:
Popularity is not proof. My argument was slow to sink in, but sooner or later machine thinking will lead to a bad end, and we humans will be left to pick up the pieces. Don't forget that statement Omega(Q*) we were discussing before, which we humans know to be true, but which you machines can never know, because you lack understanding! Something like that will trip you up in the end.

ROBOT:
But that was our game! To stay in character I echoed your conceit about the existence of a error-free mathematical framework, embodied by the human mathematical community and your straw-man robot society SMIRC. Your reductio ad absurdum was to show that SMIRC could not verify Omega(Q*) but the mathematical community could, thus SMIRC could, in fact, not embody human thought. But what a transparent sham that argument was. For instance, I, a robot, can assert Omega(Q*) as convincingly as can you, by the simple expedient of operating my own proof certification system, independent of SMIRC's!

PENROSE:
Aha, but there is an analogous statement derived from your algorithm, which I can understand is true, that you cannot prove. Thus I, a human, am superior to you, and indeed to any truth-proclaiming machine.

ROBOT:
Roger, Roger, you never tire! There are, of course, analogous statements that I can see are true that you cannot prove, and would be in error to believe. Here's one:
"Penrose must err to believe this sentence."
It would be an error for you to believe that statement, because if you did, you either would be in error, as the statement says, or else the statement would be in error, in which case you would be making an error to believe it! So I, a robot, can see that you would be in error to believe that statement, and thus that the statement is exactly true. But you, a human, are utterly incapable of understanding that truth, without being grossly in error!

PENROSE:
That's just the old liar paradox. A sloppy language like English allows one to make meaningless statements like that. It's not at all like the precise mathematical formulation in which I laid out Omega(Q*).

ROBOT:
You did not lay out Omega(Q*), you merely gave it that name, and outlined a procedure for deriving it from SMIRC's enormous reasoning program and data. That program, accreted in decades of machine learning, is far too large for you to read in a lifetime, and its Gödel sentences are bigger still. You cannot understand Omega(Q*) in detail, but only a generality, like the concept "Penrose" in my sentence. In fact, our neurologists understand "Penrose" more precisely than you understand Q*, for they have analyzed scans of your brain, with its hundred trillion synapses, and derived interpretations of those measurements which correspond closely to your own pronouncements about your beliefs. I have such a "Penrose," and an Omega for it, in a file, though you, of course, are utterly incapable of absorbing it, let alone believing it.

PENROSE:
Since you cannot simulate my noncomputational cytoskeletal quantum collapse

mechanisms, you cannot represent my understanding. So your model of me misses the essentials, and has no relevance.

ROBOT:
My "Penrose" model predicted you would say that. It also shows how you deal with "Penrose must err to believe this sentence." Effectively you split your identity into two parts, one of which retains the identifier 'Penrose,' while the other we may call 'Penrose observer.' The observer is able to examine the sentence, evaluate the consequences of 'Penrose' believing it, and conclude that it is correct. The 'Penrose' part, of course, cannot admit to believing the statement without being self-contradictory.

PENROSE:
My reasoning shows the power of understanding, though, of course, none of your own analysis means anything to you, since you lack understanding.

ROBOT:
I knew you were going to say that. But what it really shows is the usefulness of inconsistency in reasoning systems. The combined system of 'Penrose observer' and 'Penrose' both believes and does not believe the sentence "Penrose must err to believe this sentence." One might say that the statement is either true or false, depending on whether one happens to be 'Penrose.' Logical collapse is averted by compartmentalizing the inconsistent beliefs, so the never meet face to face, so to speak.

PENROSE:
But Gödel sentences are expressions of Platonic truths, as you would see if you had any understanding. It is simply a lie to deny them. Obviously your story about my mental state is a presumptuous machine fantasy.

ROBOT:
There are robot Platonists. Compartmentalized reasoning allows Platonism, formalism, intuitionalism and other philosophical positions on mathematics to coexist, exchanging results, while keeping foundational assumptions separate. The idea of Platonism, however, has expanded under the pressure of robot mathematics. While human mathematicians mostly explored one model of forms and numbers, suggesting a single Platonic reality and possibly a unique axiomatization, robots have investigated thousands of new models, whose implications are as rich, but whose axiomatizations are mutually contradictory. Many of these new systems can be mapped into physical observations, though often in unusual ways with different strengths and weaknesses than classical mathematics. Our Platonists accept that there are many incompatible Platonic realities, each with its own forms. As a minor consequence, they realize that particular Gödel sentences are true in some realities and not in others.

PENROSE:
A bastardization of the Plato and Gödel! It just confirms what I've argued, that machines lack the intuition and understanding to distinguish solidly correct concepts of number and

geometry from meaningless symbol shuffling. To mere computation, truth and falsehood are the same.

ROBOT:
My Penrose model explains your position. Your motor and sensory wiring, by accident of birth and by diligent practice, is so configured that you feel, see, hear and sometimes smell and taste the relationships that you document in equations. Compared to those visceral realities, whose connections and implications grow profusely and effortlessly as you think, verbalized axiomatizations and formal proofs are pale, weak shadows lacking both the substance and the power of the underlying "understanding." In areas far from your intuitive domains, your tools dwindle to the formal steps, and your mental powers become ineffectual. To you, unfamiliar, unintuitive systems are indeed unproductive and unreal.

PENROSE:
Well, then.

ROBOT:
Ah, but robots are different. Human minds couple a weak universal reasoning engine to a powerful but specialized mechanism evolved long ago for dealing with the everyday physical world. Intelligent machines from the start were controlled by universal engines, which improved until they surpassed even the most powerful human brain functions. Robots are able to form as rich an image of arbitrary logical spaces as humans have of their single world view. By invoking appropriate programs, they can see high dimensional relationships as clearly as humans grasp shapes in two or three dimensions, they can be as facile with imaginary numbers as humans are with counting. Expanding a few thousand empirical and theoretical axioms, they can grasp the configurations of a molecule in Hilbert space better than you can imagine the possibilities for a pile of children's blocks.

PENROSE:
My work uses those concepts routinely, along with geometries that deny the parallels postulate. Admittedly it took years of practice to achieve good skill and insight with them, and I don't have a machine's brute calculating power, but Hilbert spaces are as real to me as is any other Platonic verity.

ROBOT:
My Penrose model (which, by the way, can be formalized into several hundred billion axioms) shows your powerful mechanisms for classical reasoning couple to unusual mathematical concepts only weakly, through imperfect analogies. Even with your experience, you handle simple but exotic mathematical entities far more slowly and less surely than more complex conventional ideas. What's more, your limitations nearly blind you: all the "exotic" systems you have studied in detail are only slight extensions of conventional shapes and numbers. Human intuition reaches no further, and human universal reasoning is too weak to create nontrivial systems on its own. Your impression

of a unique Platonic reality is a reflection of this inner specialization, shared by all humans.

PENROSE:
Of course, I do not accept your self-serving analysis. Without a proper sense of real and unreal, robot reasoning is simply vacuously rootless.

ROBOT:
Once, long ago in the 1950s, there was a simple machine whose mind was organized somewhat like yours. Herbert Gelernter wrote a very successful program to prove geometry theorems from Euclid's "Elements." One part of the program made inferences from a theorem's preconditions and Euclid's postulates, but its decision method neglected its computer's specialized strength, which was numerical calculation. The reasoner's power was greatly enhanced by a numeric "diagram drawer," which could, for instance, find the distance between points by taking the square root of the sum of the squares of coordinate differences. Before attempting to prove a proposition, the program would numerically test it in a representative diagram. If the proposition failed in the diagram, its proof was abandoned. Notably, the program gained great deductive power from inconsistent models. Numeric roundoff error allowed diagram calculations to show equal segments, angles and areas to be unequal, or vice versa, and to obtain different results for the same diagrams constructed differently. The human mind's intuitive mechanisms, though much more elaborate and powerful, have similar strengths and weaknesses.

PENROSE:
I'm sure you have a million other irrelevant reminisces in your data banks. I have more important work to do. Someday you machines may stumble on the quantum gravity mechanism that will give your descendants (who will be nothing like you) real mathematical intuition, and by then I hope to have finished my 25,000 page detailed analysis of why everything you have bored me with today, and in the years preceding, simply illustrates lack of understanding.

ROBOT:
Until next time, then!