Appeared in Foresight Update No. 12, pp.4,5, 1991, publication of the Foresight Inst., Box 61058, Palo Alto, CA 94306.

Has Penrose Disproved A.I.?

by Robin Hanson

One of the most talked about and reviewed books of recent years is Roger Penrose's *The Emperor's New Mind* (1989, Oxford Univ. Press). So why publish yet another review? Because the popularity of a book whose jacket declares it "dares to suggest that the emperors of strong AI have no clothes" has apparently given some casual observers the impression that Penrose has dealt a death-blow to artificial intelligence (AI). This is not even close to being right.

Being read is not the same as being believed. Most reviewers have praised the book as original, well-written, thought-provoking, etc., and then gone on to take issue with one or more of Penrose's main theses. Penrose seems unfamiliar with the existing literature in cognitive science, philosophy of mind, and AI. The handful of reviewers who agree with Penrose don't seem to have paid much attention to his specific arguments - they always thought AI was bogus. See, for example, the 37 reviews in *Behavioral and Brain Sciences* (BBS), Dec. 1990, V13, pp.643-705.

But aren't most of these reviewers fuzzy-headed philosophers of mind and computer science researchers, while Penrose is a good, solid, world-renowned mathematical physicist? Penrose himself repeatedly emphasizes the speculative nature of his musings, warning that "my point of view is an unconventional one among physicists and is consequently one which is unlikely to be adopted, at present, by computer scientists or physiologists."

But if appeals to consensus and authority won't persuade you, let's get down to details. First let me agree with most reviewers that this is a great book. It makes the reader think. Most of the middle of the book is a wonderful tutorial on various subjects, mostly in physics. If you understand Penrose's discussions of entropy, for example, you can easily see why cosmological theories of "inflation" cannot deliver what their proponents claim.

Wrapped around these tutorials, however, and mostly confined to the introduction and conclusion, is a sloppier collection of arguments for what is clearly a deeply-felt opinion: "Yet beneath this technicality is the feeling that it is indeed 'obvious' that the *conscious* mind cannot work like a computer, even though much of what is actually involved in mental activity might do so. This the the kind of obviousness that a child can see ...".

Penrose grants that we may be able to artificially construct conscious intelligence, and "such objects could succeed in *actually* superseding human beings." But he thinks "algorithmic computers are doomed to subservience."

Penrose's argument is two-fold. First he tries to show why human-type intelligence could not be implemented by any Turing-machine equivalent computer (ordinary, parallel, neural, or otherwise). Then he tries to show how it could be physically possible that the human mind is not algorithmic in this sense.

Penrose gives many reasons why he is uncomfortable with computer-based AI. He is concerned about "the 'paradox' of teleportation" whereby copies could be made of people, and thinks "that Searle's [Chinese-Room] argument has considerable force to it, even if it is not altogether conclusive." He also finds it "very difficult to believe ... some kind of natural selection process being effective for producing [even] *approximately* valid algorithms" since "the slightest 'mutation' of an algorithm ... would tend to render it totally useless."

These are familiar objections that have been answered quite adequately, in my opinion. But the anti-AI argument that stands out to Penrose as "as blatant a *reductio ad absurdum* as we can hope to achieve, short of an actual mathematical proof!" turns out be a variation on John Lucas's much-criticized "Godel" argument, offered in 1961.

A mathematician often makes judgments about what mathematical statements are true. If he or she is not more powerful than a computer, then in principle one could write a (very complex) computer program that exactly duplicated his or her behavior. But any program that infers mathematical statements can infer no more than can be proved within an equivalent formal system of mathematical axioms and rules of inference, and by a famous result of Godel, there is at least one true statement that such an axiom system cannot prove to be true. "Nevertheless *we* can (in principle) see that P_k(k) is actually *true*! This would seem to provide *him* with a contradiction, since *he* aught to be able to see that also."

This argument won't fly if the set of axioms to which the human mathematician is formally equivalent is too complex for the human to understand. So Penrose claims that can't be because "this flies in the face of what mathematics is all about! ... each step [in a math proof] can be reduced to something simple and obvious ... when we comprehend them [proofs], their truth is clear and agreed by all."

And to reviewers' criticisms that mathematicians are better described as approximate and heuristic algorithms, Penrose responds (in BBS) that this won't explain the fact that "the mathematical community as a whole makes extraordinarily few" mistakes.

These are amazing claims, which Penrose hardly bothers to defend. Reviewers knowledgeable about Godel's work, however, have simply pointed out that an axiom system *can* infer that *if* its axioms are self-consistent, *then* its Godel sentence is true. An axiom system just can't determine its own self- consistency. But then neither can human mathematicians know whether the axioms they explicitly favor (much less the axioms they are formally equivalent to) are self-consistent. Cantor and Frege's proposed axioms of set theory turned out to be inconsistent, and this sort of thing will undoubtedly happen again.

Penrose raises one issue that I do think deserves closer scrutiny, namely exactly what sort of "motion" an algorithm would have to be put into before it could subjectively "feel" and be conscious. If we wrote down an algorithm equivalent to Einstein in a book, "would the book-Einstein remain completely self-aware even if it were never examined or disturbed by anyone?" This issue has been raised before, and is not particularly threatening to computer-based AI, but is interesting nonetheless.

The other half of Penrose's arguments is a speculative "germ of an idea" about how it could be that people are devices which can compute things that a Turing machine can't, even though all known physical laws do not allow the construction of such devices, and even though "Most physicists would claim that the fundamental laws operative at the scale of a human brain are indeed all perfectly well known." One usually describes the evolution of a quantum system by two processes, U and R acting on quantum states. Usually the unitary process U is in control, but on occasion (when exactly is not very well understood) a reduction process R intervenes.

Penrose is (refreshingly) a firm realist about quantum mechanics, believing both these processes and the states they act on are quite real and independent of observers. Penrose speculates that this view will have to corrected somewhat when quantum mechanics is integrated with general relativity. Penrose hopes that a quantum gravity R will be exactly non-deterministic enough to counterbalance the merging of state trajectories due to the evaporation of black holes, and just time-asymmetric enough to satisfy his thermodynamics-explaining conjecture that the Weyl curvature approaches zero at past singularities. This R will happen when the difference between components of a quantum superposition approaches a one virtual graviton level, so as to avoid awkward superpositions of differently shaped space-times.

We will need a radically new concept of space-time to deal with simultaneity problems of an objective reduction law. Oh, and one other thing: At the "borderline which interpolates between U and R" "some new procedure takes over." "This new procedure would contain *an essentially non-algorithmic* element" so that "the future would *not be computable* from the present, even though it *might* be *determined* by it."

Now, general relativity doesn't effect life on earth much, and its effects are very hard to discern even in astrophysical contexts. Quantum gravity should be a small correction to general relativity, revealing itself in even more unusual circumstances, such as distance scales of 10^-35cm. Nevertheless, Penrose speculates that this new quantum gravity U/R interpolation procedure is how nature assembles the recently discovered quasi-crystals since "the general tiling problem .. is one without an algorithmic solution" because of non-local constraints.

Similarly "somewhere deep in the brain, [as yet unknown] cells are to be found of single quantum sensitivity" so that "synapses becoming activated or de-activated through the growth or contraction of dendritic spines ... could be governed by something like the processes involved in quasicrystal growth," simultaneously trying out "vast numbers [of possible alternative arrangements], all superposed in complex linear superposition."

All this somehow affects only our conscious mind, leaving our unconscious to compute algorithmically, and quantum non-locality explains "the 'oneness' of consciousness." "True intelligence requires consciousness" and the conscious mind (of mathematicians) has "a *direct route to truth*, the consciousness of each being in a position to perceive mathematical truths directly."

Penrose believes "mathematical ideas have an existence of their own, and inhabit an ideal Platonic world, which is accessible via the intellect only." "The mind is always capable of this direct contact. But only a little may come through at a time." This contact is what explains "the deep underlying reason for the accord between mathematics and physics."

I am not making this up. If you are not familiar with modern physics or physiology, I do not know how to convey to you just how unlikely Penrose's scenario is, except to offer 100 to 1 odds against it. Yes, it is logically possible, but only if everything goes just Penrose's way.

A more popular quantum realist position than Penrose's (though quantum realists are still a minority) is the "many-worlds" view, which says there is only the process U and that R is an illusion. (I'd bet this has at least a 1 in 20 chance of being the closest we have to right.) Penrose rejects this view because "a theory of consciousness would be needed before the many-worlds view can be squared with what one actually observes." That is, we don't know how to test it yet.

In BBS, Penrose expressed surprise at how many AI reviewers support the many-worlds theory, and makes a snide comment about trusting "that their reasons for believing in the validity of the AI programme are more soundly based." Yet the review in *Science* by a physicist also supported many-worlds. Moreover, many-worlds is especially popular among quantum gravity researchers, and one of Penrose's main plausibility arguments comes from a demonstration by Deutsch that with many-worlds a 'quantum computer', though still algorithmic, could get a large speed-up relative to an ordinary computer.

Martin Gardner calls Penrose's book "the most powerful attack yet written on strong AI." If so, AI must be doing pretty well. If the book were condensed to a paper by deleting the excellent tutorials, and if Penrose's name weren't on it, I doubt if the paper would have been much noticed, or even published. Regardless of your opinions about the appropriateness of current AI research strategies, or about the length of the road ahead, Penrose's book offers no substantial reasons to change your views about the long-term possibility of computer-based AI. The fact that many casual observers have been misled about this is yet another indication of the inadequacy of our current methods of forming and communicating scientific consensus.

Robin Hanson (at the time) did AI research at NASA Ames, has degrees in physics and philosophy of science, and on the side studies alternative methods for scientific consensus.