The Emperor's New Mind, by Roger Penrose. Oxford University Press, Oxford, New York, Melbourne, 1989, xiii + 466 pp., $24.95. ISBN 0-19-851973-7

[This review appeared in *Bulletin of the American Mathematical Society*, Volume 23, Number 2, October 1990, pp. 606-616.]

Penrose doesn't believe that computers constructed according to presently known physical principles can be intelligent and conjectures that modifying quantum mechanics may be needed to explain intelligence. He also argues against what he calls "strong AI". Neither argument makes any reference to the 40 years of research in artificial intelligence (AI) as treated, for example, in Charniak and McDermott (1985). Nevertheless, artificial intelligence is relevant, and we'll begin with that.

The goal of AI is to understand intelligence well enough to make intelligent computer programs. It studies problems requiring intelligence for their solution and identifies and programs the intellectual mechanisms that are involved. AI has developed much more as a branch of computer science and applied mathematics than as a branch of biology. Mostly it develops, tests and makes theories about computer programs instead of making experiments and theories in psychology or neurophysiology.

The most interesting and fundamental problems of AI concern trying to make programs that can achieve goals in what we call the *common sense informatic situation*. People confront such situations in daily life and also in practicing science and mathematics. It is distinguished from the informatic situation within an already formalized theory by the following features.

1. Partial knowledge of both general phenomena and particular situations. The effect of spilling a bowl of hot soup on a table cloth is subject to laws governing absorption as well as to the equations of hydrodynamics. A computer program to predict who will jump out of the way needs facts about human motivation, human ability to observe and act as well as information about the physics. None of this information usefully takes the form of differential equations.

2. It isn't known in advance of action what phenomena have to be taken into account. We would consider stupid a person who couldn't modify his travel plan to take into account the need to stay away from a riot in an airport.

3. Even when the problem solving situation is subject to fully known laws, e.g. chess or proving theorems within an axiomatic system, computational complexity can force approximating the problem by systems whose laws are not fully known.

Faced with these problems, AI has sometimes had to retreat when the limited state of the art requires it. Simplifying assumptions are made that omit important phenomena. For example the MYCIN *expert system* for diagnosing bacterial infections of the blood knows about many symptoms and many bacteria but it doesn't know about doctors or hospitals or even processes occurring in time. This limits its utility to situations in which a human provides the common sense that takes into account what the program doesn't provide for. Other AI systems take more into account, but none today have human-level common sense knowledge or reasoning ability.

The methodology of AI involves combinations of epistemology and heuristics. Facts are represented by formulas of logic and other data structures, and programs manipulate these facts, sometimes by logical reasoning and sometimes by ad hoc devices.

Progress in AI is made by

1. Representing more kinds of general facts about the world by logical formulas or in other suitable ways.

2. Identifying intellectual mechanisms, e.g. those beyond logical deduction involved in common sense reasoning.

3. Representing the approximate concepts used by people in common sense reasoning.

4. Devising better algorithms for searching the space of possibilities, e.g. better ways of making computers do logical deduction.

Like other sciences AI gives rise to mathematical problems and suggests new mathematics. The most substantial and paradigmatic of these so far is the formalization of nonmonotonic reasoning.

All varieties of mathematical logic proposed prior to the late 1970s are monotonic in the sense that the set of conclusions is a monotonic increasing function of the set of premises. One can find many historical indications of people noticing that human reasoning is often nonmonotonic—adding a premise causes the retraction of a conclusion. It was often accompanied by the mistaken intuition that if only the language were more precise, e.g. embodied probabilities explicitly, the apparent nonmonotonicity would go away. It was consideration of how to make computers reason in common sense situations that led to pinning down and formalizing nonmonotonic reasoning

The systems for formalizing nonmonotonic reasoning in logic are of two main kinds. One, called circumscription, involves minimizing the set of tuples for which a predicate is true, subject to preserving the truth of an axiom and with certain predicate and function symbols variable and others fixed. It is a logical analogue of the calculus of variations though far less developed.

Suppose we require the extension of a predicate P to be a relative minimum, where another predicate Q is allowed to vary in achieving the minimum and a third predicate R is taken as a non-varied parameter. Suppose further that P, Q and R are required to satisfy a formula $A(P, Q, R)$. Any relative minimum P satisfies the second order formula

$$A(P, Q, R) \wedge \forall P'Q'(A(P', Q', R) \supset \neg(P' < P)),$$

where $<$ is defined by

$$P' < P \equiv \forall x(P'(x) \supset P(x)) \wedge \exists x(\neg P'(x) \wedge P(x)).$$

If $A(P, Q, R)$ is the conjunction of the facts we are taking into account, we see that circumscription is nonmonotonic, because conjoining another fact to $A(P, Q, R)$ and doing the minimization of P again can result in losing some of the consequences of the original minimization.

Here's an example. Suppose a car won't start. We have facts about the many things can go wrong with a car, and we also have facts about the present symptoms. $A(P, Q, R)$ stands for our facts and $P(x)$ stands for 'x is wrong with the car'. Circumscribing P corresponds to conjecturing that nothing more is wrong with the car than will account for the symptoms so far observed and expressed by formulas. If another symptom is observed, then doing the circumscription again may lead to new conclusions.

Applications to formalizing common sense often require minimizing several predicates in several variables with priorities among the predicates. Mathematical questions arise such as whether a minimum exists and when the above second order formula is equivalent to a first order formula.

The second kind of nonmonotonic system is based on the idea that the set of propositions that are believed has to have a certain coherence and is a fixed point of a certain operator. Ginsberg (1987) contains a selection of papers, both on the logic of nonmonotonic reasoning and on its application to formalizing common sense knowledge and reasoning.

The main difficulties in formalizing common sense are not technical mathematical problems. Rather they involve deciding on an adequately general set of predicates and functions and formulas to represent common sense knowledge. It is also necessary to decide what objects to admit to the universe such as "things that can go wrong with a car".

More innovations than nonmonotonic reasoning will be needed in logic itself, e.g. better reflexion principles and formalization of context, before computer programs will be able to match human reasoning in the common sense informatic situation. These and other conceptual problems make it possible that it will take a long time to reach human-level AI, but present progress provides reason for encouragement about achieving this goal with computer programs.

The Book

Most of the book is expository, perhaps aimed at bringing a layman to the point of understanding the author's proposals and the reasons for them. The exposition is elegant, but I think a person who has to be told about complex numbers will miss much that is essential. Topics covered include Turing machines, Penrose tiles, the Mandelbrot set, Gödel's theorem, the philosophy of mathematics, the interpretations of quantum mechanics including the Einstein-Podolsky-Rosen *Gedanken* experiment, general relativity including black holes and the prospects for a theory of quantum gravitation. Using LISP rather than Turing machines for discussing computability and Gödel's theorem would have given a shorter and more comprehensible exposition.

Before the expository part, Penrose undertakes to refute the "strong AI" thesis which was invented by the philosopher John Searle in order to be refuted. It has some relation to current opinions among artificial intelligence researchers, but it oversimplifies by ignoring AI's emphasis on knowledge and not just algorithms. As Penrose uses the term, it is the thesis that intelligence is a matter of having the right algorithm.

While Penrose thinks that a machine relying on classical physics won't ever have human-level performance, he uses some of Searle's arguments that even if it did, it wouldn't really be thinking.

Searle's (1980) "Chinese room" contains a man who knows no Chinese. He uses a book of rules to form Chinese replies to Chinese sentences passed in to him. Searle is willing to suppose that this process results in an intelligent Chinese conversation, but points out that the man performing this task doesn't understand the conversation. Likewise, Searle argues, and Penrose agrees, a machine carrying out the procedure wouldn't understand Chinese. Therefore, machines can't understand.

The best answer (published together with Searle's paper) was the "system answer". Indeed the man needn't know Chinese, but the "program" embodied in the book of rules for which the man serves as the hardware interpreter would essentially have to know Chinese in order to produce a non-trivial Chinese conversation. If the man had memorized the rules, we would have to distinguish between his personality and the Chinese personality he was interpreting.

Such situations are common in computing. A computer time-shares many programs, and some of these programs may be interpreters of programming languages or expert systems. In such a situation it is misleading to ascribe a program's capabilities to the computer, because different programs on the same computer have different capabilities. Human hardware doesn't ordinarily support multiple personalities, so using the same name for the physical person and the personality rarely leads to error.

Conducting an interesting human-level general conversation is beyond the current state of AI, although it is often possible to fool naive people as fortune tellers do. A real intelligent general conversation will require putting into the system real knowledge of the world, and the rules for manipulating it might fit into a room full of paper and might not, and the speed at which a person could look them up and interpret them might be slow by a factor of only a hundred, or it might turn out to be a million.

According to current AI ideas, besides having lots of explicitly represented knowledge, a Chinese room program will probably have to be introspective, i.e. it will have to be able to observe its memory and generate from this observation propositions about how it is doing. This will look like consciousness to an external observer just as human intelligent behavior leads to our ascribing consciousness to each other.

Penrose ignores this, saying (p. 412), "The *judgement-forming* that I am claiming is the hallmark of consciousness is *itself* something that the AI people would have no concept of how to program on a computer." In fact most of the AI literature discusses the representation of facts and judgments from them in the memory of the machine. To use AI jargon, the epistemological part of AI is as prominent as the heuristic part.

The Penrose argument against AI of most interest to mathematicians is that whatever system of axioms a computer is programmed to work in, e.g. Zermelo-Fraenkel set theory, a man can form a Gödel sentence for the system, true but not provable within the system.

The simplest reply to Penrose is that forming a Gödel sentence from a proof predicate expression is just a one line LISP program. Imagine a dialog between Penrose and a mathematics computer program.

Penrose: Tell me the logical system you use, and I'll tell you a true sentence you can't prove.

Program: You tell me what system you use, and I'll tell you a true sentence you can't prove.

Penrose: I don't use a fixed logical system.

Program: I can use any system you like, although mostly I use a system based on a variant of ZF and descended from 1980s work of David McAllester. Would you like me to print you a manual? Your proposal is like a contest to see who can name the largest number with me going first. Actually, I am prepared to accept any extension of arithmetic by the addition of self-confidence principles of the Turing-Feferman type iterated to constructive transfinite ordinals.

Penrose: But the constructive ordinals aren't recursively enumerable.

Program: So what? You supply the extension and whatever confidence I have in the ordinal notation, I'll grant to the theory. If you supply the confidence, I'll use the theory, and you can apply your confidence to the results.

[Turing adds to a system a statement of its consistency, thus getting a new system. Feferman adds an assertion that is essentially of the form $\forall n(provable P(n)) \supset \forall n P(n)$. We've left off some quotes.]

One mistaken intuition behind the widespread belief that a program can't do mathematics on a human level is the assumption that a machine must necessarily do mathematics within a single axiomatic system with a predefined interpretation.

Suppose we want a computer to prove theorems in arithmetic. We might choose a set of axioms for elementary arithmetic, put these axioms in the computer, and write a program to prove conjectured sentences from the axioms. This is often done, and Penrose's intuition applies to it. The Gödel sentence of the axiomatic system would be forever beyond the capabilities of the program. Nevertheless, since Gödel sentences are rather exotic, e.g. induction up to ϵ_0 is rarely required in mathematics, such programs operating within a fixed axiomatic system are good enough for most conventional mathematical purposes. We'd be very happy with a program that was good at proving those theorems that have proofs in Peano arithmetic. However, to get anything like the ability to look at mathematical systems from the outside, we must proceed differently.

Using a convenient set theory, e.g. ZF, axiomatize the notion of first order axiomatic theory, the notion of interpretation and the notion of a sentence holding in an interpretation. Then Gödel's theorem is just an ordinary theorem of this theory and the fact that the Gödel sentence holds in models of the axioms, if any exist, is just an ordinary theorem. Indeed the Boyer-Moore interactive theorem prover has been used by Shankar (1986) to prove Gödel's theorem, although not in this generality. See also (Quaife 1988).

Besides the ability to use formalized metamathematics a mathematician program will need to give credence to conjectures based on less than conclusive evidence, just as human mathematicians give credence to the axiom of choice. Many other mathematical, computer science and even philosophical problems will arise in such an effort.

Penrose mentions the ascription of beliefs to thermostats. I'm responsible for this (McCarthy 1979), although Penrose doesn't refer to the actual article. A thermostat is considered to have only two possible beliefs—the room is too hot or the room is too cold. The reason for including such a simple system, which can be entirely understood physically, among those to which beliefs can be ascribed is the same as the reason for including the numbers 0 and 1 in the number system. Though numbers aren't needed for studying the null set or a set with one element, including 0 and 1 makes the number system simpler. Likewise our system for ascribing beliefs and relating them to goals and actions must include simple systems that can be understood physically. Dennett (1971) introduces the "intentional stance" in which the behavior of a system is understood in terms of its goals and beliefs and a principle of rationality: *It does what it believes will achieve its goals.* Much of what we know about the behavior of many systems is intentional.

Indeed beliefs of thermostats appear in the instructions for an electric blanket: "Don't put the control on the window sill or it will think the room is colder than it is." The manufacturer presumably thought that this way of putting it would help his customers use the blanket with satisfaction.

Penrose's Positive Ideas

Penrose wants to modify quantum mechanics to make it compatible with the variable metric of general relativity. He contrasts this with the more usual proposal to modify general relativity to make it compatible with quantum mechanics.

He begins with the perennial problem of interpreting quantum mechanics physically. He prefers an interpretation using a U formalism and an R formalism. The U formalism is the Schrödinger equation and is deterministic and objective and reversible in time. The R formalism provides the theory of measurement and is probabilistic and also objective but not reversible. Penrose discusses several other interpretations.

The Bohr interpretation gives quantum measurement a subjective character, i.e. it depends on a human observer. Penrose doesn't like that, because he wants the wave function to be objective. I share his preference.

The Bohr interpretation is often moderated to allow machines as observers but remains subject to the "paradox" of Schrödinger's cat. The cat is in a sealed chamber and may or may not be poisoned by cyanide according to whether or not a radioactive disintegration takes place in a certain time interval. Should we regard the chamber as containing either a dead cat or a live cat or as having a wave function that assigns certain complex number amplitudes to dead cat states and others to live cat states?

The Everett "many worlds interpretation" considers reality to be the wave function of the whole world with the wave functions of subsystems being merely approximations by "relative wave functions". The world is considered to be splitting all the time, so there are some worlds with a dead cat and others with a live cat. Penrose doesn't like this either.

People have interpreted quantum mechanics in various ways; Penrose's point is to change it. His idea of what to change comes from thinking about quantum gravitation and especially about black holes. Penrose says that when matter enters a black hole, information is lost, and this violates Liouville's theorem about conservation of density in phase in Hamiltonian systems. This makes the system non-reversible, which he likes.

He attributes the apparent "collapse of the wave function" when an observation occurs to conventional quantum mechanics being true only at a small scale. When the scale is large enough for the curvature of space to be significant, e.g. at the scale of an observer, he expects quantum mechanics to be wrong and something like the collapse of the wave function to occur. Although Penrose gives no details, the idea already suggests a different outcome to certain experiments than quantum mechanics predicts, i.e. when an interaction is extended in space.

Quantum mechanics began in 1905 with Einstein's explanation of the photoelectric effect, in which a photon causes an electron to be emitted from a metal. If the electron is emitted from an atom, we have an instance of collapse of the wave function. Some atom is now missing an electron, and in principle an experimenter could find it, say with a scanning tunneling microscope.

However, this piece of metal also has conduction electrons, and these are not localized to atoms; the wave function of such an electron has a significant coherence length. Suppose the photon causes such an electron to be emitted. Quantum mechanics says that the emission event need not take place at a specific atomic location, and the electron's wave function after emission need not correspond to emission from a point.

In principle, this is observable. One experiment would put parallel insulating (and opaque) stripes on the metal as narrow and close together as possible with the techniques used to make integrated circuits. The electron may then not be emitted from a single gap between the stripes but from several gaps. It will then "interfere with itself", and the pattern observed on the electron detectors after many photons have emitted electrons will have interference fringes. It seems (William Spicer, personal communication) that this is a possible, though difficult, experiment.

Quantum mechanics predicts that the wave function collapses in the atomic scale photo-emission and doesn't collapse, or at least only partially collapses, at the larger scale of the coherence length of the conduction electron. Would Penrose claim that there is some scale at which this coherence could not be observed?

The book concludes by mentioning the result of Deutsch (1985) that a quantum computer might solve some problems in polynomial time that take exponential time with a conventional computer. He disagrees with Deutsch's opinion: "The intuitive explanation of these properties places an intolerable strain on all interpretations of quantum theory other than Everett's".

Nothing Penrose says indicates that he could satisfy Searle that such computer could really "think" or that it would get around Gödel's theorem. This minimal conclusion made me think of a shaggy dog story. I acknowledge the priority of Daniel Dennett, *Times Literary Supplement*, in applying this metaphor.

In the Epilog, a computer answers that it cannot understand the question when asked what it feels like to be a computer. My opinion is that some future programs will find the question meaningful and have a variety of answers based on their ability to observe the reasoning process that their programmers had to give them in order that they could do their jobs. The answers are unlikely to resemble those given by people, because it won't be advantageous to give programs the kind of motivational and emotional structure we have inherited from our ancestors.

References:

Charniak, E. and D. McDermott (1985): *Introduction to Artificial Intelligence*, Addison-Wesley.

Dennett, D.C. (1971): "Intentional Systems", *Journal of Philosophy* vol. 68, No. 4, Feb. 25., Reprinted in his *Brainstorms*, Bradford Books, 1978.

Deutsch, D. (1985): "Quantum theory, the Church-Turing principle and the universal quantum computer", Proc. R. Soc. Lond. A **400**, 97-117.

Feferman, S (1989): "Turing in the Land of $O(z)$." in *The Universal Turing Machine: A Half-Century Survey*, edited by Rolf Herken, Oxford.

Ginsberg, M. (ed.) (1987): *Readings in Nonmonotonic Reasoning*, Morgan-Kaufmann, 481 p.

McCarthy, John (1979): "Ascribing Mental Qualities to Machines" in *Philosophical Perspectives in Artificial Intelligence*, Ringle, Martin (ed.), Harvester Press, July 1979.

Quaife, A. (1988): "Automated Proofs of Löb's Theorem and Gödel's Two Incompleteness Theorems", *Journal of Automated Reasoning*, vol. 4, No. 2, pp 219-231.

Searle, John (1980): "Minds, Brains and Programs" in *Behavioral and Brain Sciences*, Vol. 3. No. 3, pp. 417-458.

Shankar, N. (1986): "Proof-checking Metamathematics", PhD Thesis, Computer Science Department, The University of Texas at Austin.

John McCarthy
Stanford University

This draft of PENROS.1[w90,JMC] TEXed on 1994 October 21 at 9:46 a.m.