Is Thinking Computable?

Peter J. Denning

$/N$-$6/$·CR

43059

$\beta\beta$

31 Jan 90

RIACS Technical Report TR-90.2

NASA Cooperative Agreement Number NCC 2-387

RIACS

Research Institute for Advanced Computer Science
An Institute of the Universities Space Research Association

Is Thinking Computable?

Peter J. Denning

Research Institute for Advanced Computer Science
NASA Ames Research Center

Strong AI claims that conscious thought can arise in computers containing the right algorithms even though none of the programs or components of those computers "understands" what is going on. As proof, it asserts that brains are finite webs of neurons, each with a definite function governed by the laws of physics; this web has a set of equations that can be solved (or simulated) by a sufficiently powerful computer. Strong AI claims the Turing test as a criterion of success. A recent debate in *Scientific American* concludes that the Turing test is not sufficient, but leaves intact the underlying premise that thought is a computable process. The recent book by Roger Penrose, however, offers a sharp challenge, arguing that the laws of quantum physics may govern mental processes and that these laws may not be computable. In every area of mathematics and physics, Penrose finds evidence of nonalgorithmic human activity and concludes that mental processes are inherently more powerful than computational processes.

Is Thinking Computable?

Peter J. Denning

Research Institute for Advanced Computer Science

31 Jan 90

The vision of thinking computers fascinates people and sells magazines and books. For decades the advocates of "strong AI" (artificial intelligence) have claimed that within one or two hundred years electronic machines will be able to do everything a human can do. They see our minds as "computers made of meat," subject to the laws of physics; as soon as we understand those laws and the physical structure of the brain, we will be able to construct computing machines that solve the "differential equations of mind" in real time and exhibit behavior exactly like ours. These machines will experience emotions, judge truth, appreciate beauty, understand, be self-conscious and intelligent, and have free wills. A few advocates go so far as to speculate that the machines will be better than we are in every way and will eventually succeed *Homo sapiens* on the evolutionary scale.

Some philosophers and scientists strongly disagree. They see computers as no different from machines of levers, wheels, moving balls, valves, or pneumatic pipes; although electronic machines can peform tasks of much greater complexity in a given

time, there is nothing essentially different about them. These skeptics see no way that any such machine could come to "understand" what it does. Indeed, they argue that "understanding" and "thinking" are meaningless concepts for machines. Expert systems are unlikely to achieve competence beyond what a "mindless, procedural bureaucracy" is capable of. Even though computers now play chess at the grand-master level, almost no one says that they have insight or an understanding of chess; they are programmed simply to perform "brute-force searches" of possible future board configurations.

I have summarized these arguments in two previous columns (1,2), and I am returning to them now because of two new important contributions to the ongoing discussion. The first is a debate in *Scientific American* between John Searle and Paul and Patricia Churchland. The other is a new book by Roger Penrose. I will discuss these works and add some reflections of my own.

In the *Scientific American* debate (3,4), Searle, a philosopher from the University of California at Berkeley, argues that no computer program can function like a mind; the Churchlands, philosophers from the University of California at San Diego, argue that systems mimicking the brain's structure can do so. The editors arranged an exchange: each side challenges the other's arguments and refutations. Neither is swayed by the other's arguments.

Both sides begin with the test Alan Turing proposed in 1950, an imitation game in which an interrogator asks questions of a human and a machine; if the interrogator is unable to distinguish the two, the machine passes the test and is declared intelligent (5). Turing replaced the question, "Can a machine think?" with "Can the interrogator

distinguish the two in an imitation game?'' because he considered the former question so imprecise as to be meaningless. His own opinion was that by the year 2000 there would exist machines capable of fooling the interrogator for at least five minutesin 30% of the games played. Turing's test is taken as a criterion of machine intelligence by advocates of strong AI.

Searle reviews his own Chinese Room argument, in which a man who understands no Chinese translates between incoming messages in Chinese and outgoing messages in Chinese by performing pattern replacements following rules in a book. According to Chinese observers on the outside, the room passes the Turing test by conversing in Chinese; but the man in the room has absolutely no understanding of what is going on. Searle maintains that a computer is no different: any machine that might pass the Turing test cannot be said to be thinking. Human brains have the capacity, conferred by their specific biology, to attach meanings to symbols, a fact that differentiates them from computer programs. Simulation is not the same as duplication, and so Searle wonders why so many are prepared to accept a simulation of thinking as actual thought when they would not do the same for a computer simulation of digestion.

The Churchlands agree that the Turing test is not a sufficient condition for conscious intelligence. But they reject Searle's claim that an algorithm cannot be intelligent in principle. They argue that a brain is a finite, complicated web of neurons, each of which performs a definite function governed by the laws of physics. A set of mathematical equations relates all the signals appearing in the web; a sufficiently powerful computer would be able to solve for (or simulate) what a given brain does by solving those equations. The Churchlands recognize that the required computational

power is likely to be achieved only within the architecture of neural networks that mimic the structure of the brain. In such systems, intelligent behavior arises macroscopically, from the collective effects of simple neuron firings, and thus the individual neurons do not need to "understand" anything.

I found it fascinating that both sides presented coherent interpretations of strong AI -- with conflicting conclusions. Each side is sure it is "right" and is impervious to the other's counterarguments. None of the theories of machine intelligence I am aware of addresses this all-too-human phenomenon.

It is also interesting that both sides, following the tradition begun by Turing, dismiss the question "What is thinking?" as meaningless. But this question remains at the heart of the debate. You will see shortly that it is central to Penrose's investigation.

What we think thinking is has been a moving target throughout history. For two hundred years under the ascendancy of Newtonian mechanics beginning in the early 1700s, everyone accepted the universe as a marvelous clockwork system governed by a few simple laws. In this tradition the epitomy of human thought was problem-solving through logical deduction, man's path to exploring God's universe. The quest for a complete understanding of thought led to attempts beginning in the 1800s to formulate a universal system of logic in which all statements could be mechanically checked for validity. But the hope for such a system was dashed in the 1930s by the incompleteness theorem of Gödel and the incomputability theorem of Turing. Still, the idea that thinking was somehow a mechanical process lived on in Turing and guided his formulation of testing for intelligence. The idea of a computer thinking didn't seem the slightest bit strange to him.

Today a different interpretation of thinking is challenging the old idea. Many of us believe that thinking is not logical deduction, but the creation of new ideas. Logical deduction seems too mechanical. When we recall our moments of insight, we often say that our emotional state affected us and that we had a bodily sense of our creation before we could put it into words. We regard thinking as a phenomenon that occurs before articulation in language, and it seems that machines, which are programmed inside language, cannot generate actions outside language.

I have no doubt that fifty years from now there will be many machines that perform tasks that today we associate with thinking -- and that people will still regard them as only machines. Interpretations of thinking will have shifted farther, preserving a clear distinction between human and machine.

I turn now to Penrose's book, *The Emperor's New Mind (6)*. Penrose, a mathematician and physicist at Oxford University, mounts the most serious attack on strong AI I that have yet seen. This is not an easy book: Penrose leaves few stones unturned as he considers a broad range of speculations about mind, consciousness, and thinking. He takes his readers on an odyssey through a heady array of topics, including algorithms, Turing machines, Mandelbrot sets, formal systems, undecidability, incompleteness theorems, nonrecursive sets, Newtonian mechanics, space-time, phase spaces, relativity, quantum mechanics, entropy, cosmology, black holes, quantum gravity, brains, neurophysiology, animal consciousness, and more. In each topic he finds abundant evidence of human actions that are not algorithmic, concluding with the claim that a full understanding of mind awaits the development of quantum theories of physics as yet unknown to us.

Penrose agrees with Searle that the Turing test is an inadequate description of intelligence, but he challenges Searle's assumption that computers might pass the test. He asserts repeatedly that mental processes are inherently more powerful than computational processes. He points to the principle of universal computation -- the idea that a general-purpose computer can simulate any other machine -- as the basis for the widespread belief that algorithms must be the essence of thought. As a consequence, Penrose devotes considerable attention to the subject of noncomputable functions, such as the halting problem (is there a program which any given algorithm halts for a given input?), and he returns frequently to the idea that most of the questions about science that we consider interesting are not solvable by any general algorithm. Minds are constantly coming up with solutions to questions for which there is no general algorithm. How, he asks, could an algorithm have discovered theorems like Turing's and Gödel's that tell us what algorithms cannot do?

Penrose next takes on the strong-AI claim that we will one day have a sufficient understanding of the laws of physics and the structure of the brain to conduct an exact simulation by computer. What is physics? he asks. Is physics capable of complete understanding? What is an exact simulation? After exploring the failures of Newtonian physics that led to the formulation of relativity theory and then quantum mechanics, Penrose argues that the laws of physics at the quantum level may be determinate but not computable. Because some mental phenomena operate at scales where quantum effects may exert an influence, the functions representing the mind may not be computable, and thus an exact mechanical simulation may not be possible.

Although these suggestions are not provable given our current state of knowledge, Penrose has nonetheless offered a sharp metaphysical challenge to strong AI. The presupposition of a definite set of computable mathematical equations that determine a thinking being's next response begs the question because it implicitly assumes that all mental processes are algorithmic. If, as Penrose suggests, important physical processes of the brain are not computable, then computable equations would be only an approximation; they would leave out the quantum effects on which the conscious thought of the brain may depend.

Penrose does not, in my view, deal adequately with the shifting interpretations of consciousness and thinking. It is precisely the motion of these moving targets that must be dealt with. Penrose holds that consciousness has something to do with awareness of timeless Platonic realities: "When mathematicians communicate, [mathematical understanding] is made possible by each one having a *direct route to truth*, the consciousness of each being in a position to perceive mathematical truths directly" (p. 428).

I have found the biological interpretation of self and consciousness of Humberto Maturana and Francesco Varela (7) to be a good corrective to the narrow view expressed by Penrose. Maturana and Varela say that consciousness is associated with (but not uniquely determined by) the way we observe things. There are levels of consciousness, ranging from responding reflexively and following rules mindlessly to observing oneself as an observer. Each observer operates within a system of interpretation that includes biases, prejudices, presuppositions, culture, history, and values and that affects what can be seen or not seen, what is important or not important, and what is held as true or not

true. As conscious beings, we must constantly reckon with different observers of the same phenomena. For example, Searle and the Churchlands are different observers of strong AI and have reached different conclusions from their observations of the same phenomena.

The invention of interpretations is a fundamentally human activity that is intimately involved with our understanding of truth. As scientists, we like to say that scientific laws and mathematical theorems already exist awaiting discovery. But if we carefully examine the processes of science, we find paradigms other than discovery. Roald Hoffmann says that creation of new substances not found in nature is the dominant activity in disciplines such as chemistry and molecular biology (8). Bruno Latour goes farther, observing that in practice a statement is accepted as true by a community if no one has been able to produce evidence or an argument that persuades others to dissent (9). Science is a process of constructing facts, and different scientific communities can construct different systems of interpretation of the same physical phenomena. Western and Eastern medicine, for example, are two scientifically valid systems of interpretation about disease and human disorders; each recommends different interventions for the same symptoms and sees phenomena that are invisible to the other, and their interpretations are not easily reconciled.

Considerations such as about the variousness of truth and make it difficult for me to accept Penrose's speculations about links between consciousness and Platonic truth. For me, the existence of multiple, incomplete interpretations actually supports Penrose's basic claims about mental as opposed to computational phenomena. Like a system of logic, an interpretation cannot include all phenomena. Our powers of conscious

observation give us a capacity to step outside a particular interpretation and devise extensions or alternatives. Thus consciousness itself cannot be captured by any fixed description or interpretation. How then can consciousness be captured by an algorithm, which by its very nature is a fixed interpretation? This question applies also to algorithms that are apparently designed to shift their interpretations, because the rules for shifting constitute an interpretation themselves.

Although Penrose has left us with a great many questions that will occupy philosophers among us for many years, it is well to remember that we will continue to build practical systems that take on increasingly sophisticated tasks, such as recognition of speech and visual shapes, diagnosis, advising, symbolic mathematics, and robotics.

We humans see ourselves at the top of the current evolutionary scale. Although it intrigues us that we might have a godlike power to create beings more advanced than ourselves, we are also threatened by that possibility. Searle, the Churchlands, and Penrose have bolstered our confidence in the belief that we are more than mechanical devices. We can rest a little easier, always keeping a watchful eye on the literature in case someone comes up with a plausible argument that machines may, one day, think.

References

1.   P. J. Denning. 1986. "Will Machines Ever Think?" *American Scientist 74*, 4 (July-August). 344-346.

2.   P. J. Denning. 1988. "Blindness in the Design of Intelligent Systems," *American Scientist 76*, 2 (March-April). 118-120.

3.   J. R. Searle. 1990. "Is the Brain's mind a computer program?" *Scientific American 262*, 1 (January). 26-31.

4.   P. M. Churchland and P. S. Churchland. 1990. "Could a machine think?" *Scientific American 262*, 1 (January). 32- 37.

5.   A. M. Turing. 1950. "Computing machinery and intelligence." *Mind* Vol. LIX, No. 236. Reprinted in *The Mind's I* by D. R. Hofstadter and D. C. Dennett, Basic Books, 1981.

6.   R. Penrose. 1989. *The Emperor's New Mind.* Oxford University Press.

7.   H. Maturana and F. Varela. 1987. *The Tree of Knowledge.* Shambhala New Science Library.

8.   R. Hoffmann. 1990. "Creation and discovery." *American Scientist 78*, 1 (January-February). 14-15.

9.   B. Latour. 1987. *Science in Action.* Harvard University Press.

RIACS

Mail Stop 230-5
NASA Ames Research Center
Moffett Field, CA 94035
(415) 694-6363

The Research Institute for Advanced Computer Science
is an Institute of the
Universities Space Research Association
The American City Building
Suite 212
Columbia, MD 21044
(301) 730-2656