The New York Review of Books

VOLUME 29, NUMBER 7 · APRIL 29, 1982

# The Myth of the Computer

**By John R. Searle**

*The Mind's I: Fantasies and Reflections on Self and Soul*
composed and arranged by Douglas R. Hofstadter, by Daniel C. Dennett
Basic Books, 501 pp., $16.95

Our ordinary ways of talking about ourselves and other people, of justifying our behavior and explaining that of others, express a certain conception of human life that is so close to us, so much a part of common sense that we can hardly see it. It is a conception according to which each person has (or perhaps *is*) a mind; the contents of the mind—beliefs, fears, hopes, motives, desires, etc.—cause and therefore explain our actions; and the continuity of our minds is the source of our individuality and identity as persons.

In the past couple of centuries we have also become convinced that this common-sense psychology is grounded in the brain, that these mental states and events are somehow, we are not quite sure how, going on in the neurophysiological processes of the brain. So this leaves us with two levels at which we can describe and explain human beings: a level of common-sense psychology, which seems to work well enough in practice but which is not scientific; and a level of neurophysiology, which is certainly scientific but which even the most advanced specialists know very little about.

But couldn't there be a third possibility, a science of human beings that was not introspective common-sense psychology but was not neurophysiology either? This has been the great dream of the human sciences in the twentieth century, but so far all of the efforts have been, in varying degrees, failures. The most spectacular failure was behaviorism, but in my intellectual lifetime I have lived through exaggerated hopes placed on and disappointed by games theory, cybernetics, information theory, generative grammar, structuralism, and Freudian psychology, among others. Indeed it has become something of a scandal of twentieth-century intellectual life that we lack a science of the human mind and human behavior, that the methods of the natural sciences have produced such meager results when applied to human beings.

The latest candidate or family of candidates to fill the gap is called cognitive science, a collection of related investigations into the human mind involving psychology, philosophy, linguistics, anthropology, and artificial intelligence. Cognitive science is really the name of a family of research projects and not a theory, but many of its practitioners think that the heart of cognitive science is a theory of the mind based on artificial intelligence (AI). According to this theory minds just are computer programs of certain kinds. The main ideological aim of Hofstadter and Dennett's book is to advance this theory.

The book consists of twenty-seven essays "composed and arranged" by the two editors, all but one of them followed by "reflections" written by the composers and arrangers. The book's aim, they tell us, is "to provoke, disturb, and befuddle its readers, to make the obvious strange and, perhaps, to make the strange obvious." The result is a very heterogeneous collection, almost a hodgepodge, in which such well-known works as "Computing Machinery and Intelligence" by the British logician Alan Turing and Thomas Nagel's "What Is It Like to Be a Bat?" appear along with two pieces of fiction by Borges, and several works of science fiction—three of them by the Polish author Stanislaw Lem; and a selection by the biologist Richard Dawkins occurs alongside fantasy dialogues by Hofstadter and an imaginative piece about brain-splitting by Arnold Zuboff. In addition to their reflections, the editors include three pieces by Hofstadter and one by Dennett.

The table of contents is heavy with philosophers—Nagel, Nozick, Cherniak, Dennett, Leiber, and Smullyan, as well as the present reviewer. This is perhaps not surprising because, though the argument of the book is in favor of cognitive science, the issues it raises have mostly to do with the philosophy of cognitive science and not with actual contemporary practice.

A standard device used throughout the book is the "*Gedankenexperiment*" where we are asked to imagine some more or less fantastic eventuality as a way of challenging and testing our common sense and theoretical convictions: What if...your brain was transferred to another body, your brain was split in two, you were a brain in a vat, the information in your brain was put into some other medium; what if your brain were separated from your body but still controlled it, what if inside your head were tiny conscious demons, etc.?

With the exception of a few of the pieces, the general tone of the book is whimsical, almost playful, and the composers and arrangers are inordinately fond of puns. I think many readers will find all the conscientious whimsy, the eager cuteness, a bit wearying. One can put up with "Is the soul greater than the hum of its parts?" but one senses a certain straining for effect when one reads "Prelude...Ant Fugue" as the title of a dialogue involving an anteater. It is the sort of collection that publishers describe as "delightful."

The editors are also a bit coy about stating their own views. Their thesis is often insinuated by asking rhetorical questions. Thus we are asked "Is mentality like milk or like a song?" The correct answer is supposed to be "song." But in spite of their protestations about wanting only to "provoke, disturb, and befuddle" the editors are in fact using the book throughout as a platform to promote their own theory. There is nothing reprehensible about using your book to state your theory, but there is at least a hint of the disingenuous in advertising your book as a collection of ideologically diverse musings and then loading it largely to favor one side. They indirectly admit that this is what is going on, in the introduction, where they state the theory and discuss how they will deal with the various "roadblocks" that it faces.

The theory, which is fairly widely held in cognitive science, can be summarized in three propositions.

1. *Mind as Program*. What we call minds are simply very complex digital computer programs. Mental states are simply computer states and mental processes are computational processes. Any system whatever that had the right program, with the right input and output, would have to have mental states and processes in the same literal sense that you and I do, because that is all there is to mental states and processes, that is all that you and I have. The programs in question are "self-updating" or "self-designing" "systems of representations."

2. *The Irrelevance of the Neurophysiology of the Brain*. In the study of the mind actual biological facts about actual human and animal brains are irrelevant because the mind is an "abstract sort of thing" and human brains just happen to be among the indefinitely large number of kinds of computers that can have minds. Our minds happen to be embodied in our brains, but there is no essential connection between the mind and the brain. Any other computer with the right program would also have a mind.

Theses 1 and 2 are summarized in the introduction where the authors speak of "the emerging view of the mind as software or program—as an abstract sort of thing whose identity is independent of any particular physical embodiment."

3. *The Turing Test as the Criterion of the Mental*. The conclusive proof of the presence of mental states and capacities is the ability of a system to pass the Turing test, the test devised by Alan Turing and described in his article in this book. If a system can convince a competent expert that it has mental states then it really has those mental states. If, for example, a machine could "converse" with a native Chinese speaker in such a way as to convince the speaker that it understood Chinese then it would literally understand Chinese.

The three theses are neatly lumped together when one of the editors writes, "Minds exist in brains and may come to exist in programmed machines. If and when such machines come about, their causal powers will derive not from the substances they are made of, but from their design and the programs that run in them. And the way we will

know they have those causal powers is by talking to them and listening carefully to what they have to say."

We might call this collection of theses "strong artificial intelligence" (strong AI).[1] These theses are certainly not obviously true and they are seldom explicitly stated and defended.

Let us inquire first into how plausible it is to suppose that specific biochemical powers of the brain are really irrelevant to the mind. It is an amazing fact, by the way, that in twenty-seven pieces about the mind the editors have not seen fit to include any whose primary aim is to tell us how the brain actually works, and this omission obviously derives from their conviction that since "mind is an abstract sort of thing" the specific neurophysiology of the brain is incidental. This idea derives part of its appeal from the editors' keeping their discussion at a very abstract general level about "consciousness" and "mind" and "soul," but if you consider specific mental states and processes—being thirsty, wanting to go to the bathroom, worrying about your income tax, trying to solve math puzzles, feeling depressed, recalling the French word for "butterfly"—then it seems at least a little odd to think that the brain is so irrelevant.

Take thirst, where we actually know a little bit about how it works. Kidney secretions of renin synthesize a substance called angiotensin. This substance goes into the hypothalamus and triggers a series of neuron firings. As far as we know these neuron firings are a very large part of the cause of thirst. Now obviously there is more to be said, for example about the relations of the hypothalamic responses to the rest of the brain, about other things going on in the hypothalamus, and about the possible distinctions between the *feeling* of thirst and the *urge* to drink. Let us suppose we have filled out the story with the rest of the biochemical causal account of thirst.

Now the theses of the mind as program and the irrelevance of the brain would tell us that what matters about this story is not the specific biochemical properties of the angiotensin or the hypothalamus but only the formal computer programs that the whole sequence instantiates. Well, let's try that out as a hypothesis and see how it works. A computer can simulate the formal properties of the sequence of chemical and electrical phenomena in the production of thirst just as much as it can simulate the formal properties of anything else—we can simulate thirst just as we can simulate hurricanes, rainstorms, five-alarms fires, internal combustion engines, photosynthesis, lactation, or the flow of currency in a depressed economy. But no one in his right mind thinks that a computer simulation of a five-alarm fire will burn down the neighborhood, or that a computer simulation of an internal combustion engine will power a car or that computer simulations of lactation and photosynthesis will produce milk and sugar. To my amazement, however, I have found that a large number of people suppose that computer simulations of mental phenomena, whether at the level of brain processes or not, literally produce mental phenomena.

Again, let's try it out. Let's program our favorite PDP-10 computer with the formal program that simulates thirst. We can even program it to print out at the end "Boy, am I thirsty!" or "Won't someone please give me a drink?" etc. Now would anyone suppose that we thereby have even the slightest reason to suppose that the computer is literally thirsty? Or that any simulation of any other mental phenomena, such as understanding stories, feeling depressed, or worrying about itemized deductions, must therefore produce the real thing? The answer, alas, is that a large number of people are committed to an ideology that requires them to believe just that. So let us carry the story a step further.

The PDP-10 is powered by electricity and perhaps its electrical properties can reproduce some of the actual causal powers of the electrochemical features of the brain in producing mental states. We certainly couldn't rule out that eventuality a priori. But remember: the thesis of strong AI is that the mind is "independent of *any* particular embodiment" because the mind is just a program and the program can be run on a computer made of anything whatever provided it is stable enough and complex enough to carry the program. The actual physical computer could be an ant colony (one of their examples), a collection of beer cans, streams of toilet paper with small stones placed on the squares, men sitting on high stools with green eye shades—anything you like.

So let us imagine our thirst-simulating program running on a computer made entirely of old beer cans, millions (or billions) of old beer cans that are rigged up to levers and powered by windmills. We can imagine that the program simulates the neuron firings at the synapses by having beer cans bang into each other, thus achieving a strict correspondence between neuron firings and beer-can bangings. And at the end of the sequence a beer can pops up on which is written "I am thirsty." Now, to repeat the question, does anyone suppose that this Rube Goldberg apparatus is literally thirsty in the sense in which you and I are?

Notice that the thesis of Hofstadter and Dennett is not that *for all we know* the collection of beer cans might be thirsty but rather that if it has the right program with the right input and output it *must be* thirsty (or understand Proust or worry about its income tax or have any other mental state) because that is all the mind is, a certain kind of computer program, and any computer made of anything at all running the right program would have to have the appropriate mental states.

I believe that everything we have learned about human and animal biology suggests that what we call "mental" phenomena are as much a part of our biological natural history as any other biological phenomena, as much a part of biology as digestion, lactation, or the secretion of bile. Much of the implausibility of the strong AI thesis derives from its resolute opposition to biology; the mind is not a concrete biological phenomenon but "an abstract sort of thing."

Still, in calling attention to the implausibility of supposing that the specific casual powers of brains are irrelevant to minds I have not yet fully exposed the preposterousness of the strong AI position, held by Hofstadter and Dennett, so let us press on and examine a bit more closely the thesis of mind as program.

**D**igital computer programs by definition consist of sets of purely formal operations on formally specified symbols. The ideal computer does such things as print a 0 on the tape, move one square to the left, erase a 1, move back to the right, etc. It is common to describe this as "symbol manipulation" or, to use the term favored by Hofstadter and Dennett, the whole system is a "self-updating representational system"; but these terms are at least a bit misleading since as far as the computer is concerned the symbols don't *symbolize* anything or *represent* anything. They are just formal counters.

The computer attaches no meaning, interpretation, or content to the formal symbols; and qua computer it couldn't, because if we tried to give the computer an interpretation of its symbols we could only give it more uninterpreted symbols. The interpretation of the symbols is entirely up to the programmers and users of the computers. For example, on my pocket calculator if I print "3 x 3 = ," the calculator will print "9" but it has no idea that "3" means 3 or that "9" means 9 or that anything means anything. We might put this point by saying that the computer has a syntax but no semantics. The computer manipulates formal symbols but attaches no meaning to them, and this simple observation will enable us to refute the thesis of mind as program.

Suppose that we write a computer program to simulate the understanding of Chinese so that, for example, if the computer is asked questions in Chinese the program enables it to give answers in Chinese; if asked to summarize stories in Chinese it can give such summaries; if asked questions about the stories it has been given it will answer such questions.

Now suppose that I, who understand no Chinese at all and can't even distinguish Chinese symbols from some other kinds of symbols, am locked in a room with a number of cardboard boxes full of Chinese symbols. Suppose that I am given a book of rules in English that instruct me how to match these Chinese symbols with each other. The rules say such things as that the "squiggle-squiggle" sign is to be followed by the "squoggle-squoggle" sign. Suppose that people outside the room pass in more Chinese symbols and that following the instructions in the book I pass Chinese symbols back to them. Suppose that unknown to me the people who pass me the symbols call them "questions," and the book of instructions that I work from they call "the program"; the symbols I give back to them they call "answers to the questions" and me they call "the computer." Suppose that after a while the programmers get so good at writing the programs and I get so good at manipulating the symbols that my

answers are indistinguishable from those of native Chinese speakers. I can pass the Turing test for understanding Chinese. But all the same I still don't understand a word of Chinese and neither does any other digital computer because all the computer has is what I have: a formal program that attaches no meaning, interpretation, or content to any of the symbols.

What this simple argument shows is that no formal program by itself is sufficient for understanding, because it would always be possible in principle for an agent to go through the steps in the program and still not have the relevant understanding. And what works for Chinese would also work for other mental phenomena. I could, for example, go through the steps of the thirst-simulating program without feeling thirsty. The argument also, *en passant*, refutes the Turing test because it shows that a system, namely me, could pass the Turing test without having the appropriate mental states.[2]

The mental gymnastics that partisans of strong AI have performed in their attempts to refute this rather simple argument are truly extraordinary[3] but the method employed by Hofstadter and Dennett is more direct: they simply fabricate a direct quotation that was never uttered by me and then attack the quotation. They claim "it is a mistake to try to impute the understanding to the (incidentally) animate simulator; rather it belongs to the system as a whole, which includes what Searle casually characterizes as 'a few slips of paper.' This offhand comment, we feel, reveals how Searle's image has blinded him to the realities of the situation." The trouble is, no such "offhand comment" was ever made by me, casually or otherwise; it is a complete fabrication, as any reader can verify from my article, which they reprint in full. Moreover, it is not an understandable slip since it runs dead opposite to what I do in fact say. Furthermore it is not an idle misquotation since it is repeated no less than five times by Hofstadter and Dennett and is made the basis of their "argument."[4] Every author expects to be misinterpreted but to have a direct quotation fabricated and then made the basis of a misinterpretation is unique in my experience. Since Hofstadter and Dennett devote more space to attacking the Chinese room argument than to any other of their reflections in the book it is unfortunate that they did not trouble to get the elementary facts right.

The rest of what they have to say is mostly a repetition of points made by other authors and already answered by me. Specifically, they endorse the "systems reply" to the Chinese room argument, according to which the man in the room does not understand Chinese, but the system of which he is a part—including the instruction book, the Chinese symbols, etc.—really does understand Chinese. Adherents of this view believe, to my constant amazement, that though the man fails to understand, the *room* does understand Chinese. The obvious objection to this is that the system has no way of attaching meaning to the uninterpreted Chinese symbols, any more than the man did in the first place. The system, like the man, has a syntax but no semantics. And you can see this by simply imagining that the man internalizes the whole system. Suppose he has a super memory and a super intelligence so that he memorizes the instruction book and does all the calculations in his head. To get rid of the room, we can even suppose he works outdoors. Now since the man doesn't understand Chinese, and since there's nothing in the system that is not in the man, there is no way the system could understand Chinese. As near as I can tell Hofstadter and Dennett's only reply to this is to observe that no normal human being could perform such a feat of memory. This is of course quite true, but also quite irrelevant to the point, which, to repeat, is that from syntax alone you don't get semantics.

For reasons that seem to me utterly confused they think that my reply to one of the thought experiments actually commits me to accepting the systems reply. Suppose that the neuronal connections of a Chinese-speaking woman are broken, but suppose that a tiny, lightning-fast demon in her head makes all the connections in just the right order. Would she then understand Chinese? Assuming that the powers of her brain are fully restored the answer seems to me obviously yes. But to say that is in no way to endorse the systems reply, since in this case we are dealing with the specific causal powers of the human brain, whereas the systems reply claims that a system made of any substance at all could have mental states.

The details of how the brain works are immensely complicated and largely unknown, but some of the general

principles of the relations between brain functioning and computer programs can be stated quite simply. First, we know that brain processes cause mental phenomena. Mental states are caused by and realized in the structure of the brain. From this it follows that any system that produced mental states would have to have powers equivalent to those of the brain. Such a system might use a different chemistry, but whatever its chemistry it would have to be able to cause what the brain causes. We know from the Chinese room argument that digital computer programs by themselves are never sufficient to produce mental states. Now since brains do produce minds, and since programs by themselves can't produce minds, it follows that the way the brain does it can't be by simply instantiating a computer program. (Everything, by the way, instantiates some program or other, and brains are no exception. So in that trivial sense brains, like everything else, are digital computers.) And it also follows that if you wanted to build a machine to produce mental states, a thinking machine, you couldn't do it solely in virtue of the fact that your machine ran a certain kind of computer program. The thinking machine couldn't work solely in virtue of being a digital computer but would have to duplicate the specific causal powers of the brain.

A lot of the nonsense talked about computers nowadays stems from their relative rarity and hence mystery. As computers and robots become more common, as common as telephones, washing machines, and forklift trucks, it seems likely that this aura will disappear and people will take computers for what they are, namely useful machines. In the meantime one has to try to avoid certain recurring mistakes that keep cropping up in Hofstadter and Dennett's book as well as in other current discussions.

The first is the idea that somehow computer achievements pose some sort of threat or challenge to human beings. But the fact, for example, that a calculator can outperform even the best mathematician is no more significant or threatening than the fact that a steam shovel can outperform the best human digger. (An oddity of artificial intelligence, by the way, is the slowness of the programmers in devising a program that can beat the very best chess players. From the point of view of games theory, chess is a trivial game since each side has perfect information about the other's position and possible moves, and one has to assume that computer programs will soon be able to outperform any human chess player.)

A second fallacy is the idea that there might be some special human experience beyond computer simulation because of its special humanity. We are sometimes told that computers couldn't simulate feeling depressed or falling in love or having a sense of humor. But as far as simulation is concerned you can program your computer to print out "I am depressed," "I love Sally," or "Ha, ha," as easily as you can program it to print out "3 x 3 = 9." The real mistake is to suppose that simulation is duplication, and that mistake is the same regardless of what mental states we are talking about. A third mistake, basic to all the others, is the idea that if a computer can simulate having a certain mental state then we have the same grounds for supposing it really has that mental state as we have for supposing that human beings have that state. But we know from the Chinese room argument as well as from biology that this simple-minded behaviorism of the Turing test is mistaken.

Until computers and robots become as common as cars and until people are able to program and use them as easily as they now drive cars we are likely to continue to suffer from a certain mythological conception of digital computers. This book is very much a part of the present mythological era of the computer.[5]

### Notes

[1] "Strong" to distinguish the position from "weak" or "cautious" AI, which holds that the computer is simply a very useful tool in the study of the mind, not that the appropriately programmed computer literally has a mind.

[2] The "Chinese room argument" is stated in detail in my article "Minds, Brains, and Programs," pages 353-373 of *The Mind's I*. It originally appeared in *The Behavioral and Brain Sciences*, Vol. 3 (Cambridge University Press, 1980), along with twenty-seven responses and a reply to the responses.

[3] See the twenty-seven commentaries in *The Behavioral and Brain Sciences*, op. cit., for some examples.

[4] The initials at the end of the "Reflections" suggest that the fabrication was done by Hofstadter and not Dennett.

[5] For more balanced presentations of cognitive science see *Perspectives on Cognitive Science*, edited by Donald Norman (Ablex, Norwood, NJ, 1981); *Mind Design: Philosophy, Psychology and Artificial Intelligence*, edited by John Haugeland (Branford/MIT Press, 1981); Hubert Dreyfus, *What Computers Can't Do: A Critique of Artificial Intelligence* (Harper and Row, 1972).

**Letters**

June 24, 1982: Daniel C. Dennett, **The Myth of the Computer: An Exchange**