This work originally appeared in:

Dennett, Daniel C. "The Role of the Computer Metaphor in Understanding the Mind." In Computer Culture: the Scientific, Intellectual, and Social Impact of the Computer, edited by Heinz R. Pagels, 266-75. New York: New York Academy of Sciences, 1984.

The Role of the Computer Metaphor in Understanding the Mind

Daniel C. Dennett
Department of Philosophy
Tufts University
Medford, Massachusetts 02155

Professor Sherry Turkle has mentioned that when children look inside a computer toy, they find something utterly inscrutable: a little chip with no moving parts. They cannot make any sense of it at all. The same thing is true, of course, if you take off the top of somebody's skull and look at his brain. Absolute inscrutability. And oddly enough, it does not help if you take out your microscope and look at the details of the brain very closely. You will no more see a thought or an idea or a pain or an intention if you look at the synapses or neurotransmitters than if you look at the hypothalamus or occipital cortex or the other large parts of the brain.

There are several responses to this inscrutability or opacity of the brain. The first, and traditional, response is dualism: the ball of stuff we see between the ears could not possibly explain the mind, so the mind must be made of some other stuff altogether, some God-like, nonmechanical stuff. This is a well-known scientific dead end; in fact, it is giving up on science altogether. It amounts to "Let God do it."

Another response that I will say more about shortly is what I call mysticism about *the organic brain:* dualism is false; the mind must be the brain somehow, but it must be essentially mysterious. "I wonder if we will ever understand how!"

The third response is "Roll up your sleeves and dig in ." The brain is mysterious, in fact quite inscrutable, but let us just start at the periphery and work our way slowly in, seeing if we can make sense of it. This is often called "bottom-up" as opposed to "top-down" research in the sciences of the mind. "Top down" starts at the mind and mental events and works down, hoping to get someday to the synapses; "bottom up" starts at the synapses and hopes to work up eventually to the mind. This is a responsible and legitimate reaction to inscrutability. We have seen at this conference several good examples of research conducted in this spirit, but it is not the approach I am going to talk about here.

I am going to talk about yet a fourth reaction: the strategy of *theoretical idealization.* As John Searle put it in the panel discussion [Part VII], according to this research strategy, "the brain does not matter" - for the moment! According to this strategy, we should ignore the messy, fine details of the brain for awhile and see if we can find some theoretical idealization that will enable us to begin to get a grip on how the activities and processes of the mind might be organized.

Perhaps the reason I favor this approach is that it is the traditional philosophical approach-and I am a philosopher. The brain does not matter in traditional epistemology: you simply posit a mind, a thinking thing or "res cogitans" as Descartes put it, and then you start theorizing in an a priori fashion about the features and properties of such a knowing subject. Now in traditional epistemology (and I mean by traditional epistemology almost everything before the day before yesterday) this idealization was an *extreme* idealization. Typically the mind was supposed to be infinite or as good as infinite. In Descartes' terms it was explicitly infinite. In the terms of the logical positivists in this century it might just as well have been infinite, since no one was interested in any particular limits there might be on any actual capacities of the mind. But another curious feature of this traditional philosophical idealization of the mind is that the knowing subject was imagined to be a sort of mandarin. a person with no cares in the world, waited on hand and foot, apparently, whose only task was to avoid error at all costs. No difficulties intervened; there was never any time pressure; the goal was not ever to make a mistake. So the theories of the mind that emerged were all designed to describe methods that would permit one *always* to go from certainty to certainty, and never risk error.

The philosopher Clark Glymour recently said to me that artificial intelligence (AI) is really just "logical positivism carried on by other means." There is a certain amount of truth to that, but as we considered the matter further, several major differences became clear between traditional logical positivism and AI, which Glymour calls "android epistemology"-a good term. I think.

In android epistemology the original overidealized philosophical model is enriched by three constraints:

1. *Mechanism*. Whatever theory is proposed. one must be able to describe (perhaps with a modicum of hand waving) how in principle it could be "realized" in a mechanism. This indirect but important constraint is of course a bulwark against dualism.
2. Finitude. The proposed models of mind must all suppose the mind to be finite. to have limited resources.
3. Time pressure. The model must be able to find the "right" answers in the real time available in the real world. Life rushes on; the world will not wait for the thinking thing to ponder all possible avenues; it must be able to act intelligently under time pressure. (Professor Rabin. in his talk. brought out in a different way the importance of this constraint. This is a deep. fundamental constraint on any model of an epistemological subject or knower of the world.)

How important are these constraints? Abstract as they are. they make the game of coming up with a top-down theory of the mind deliciously difficult without being impossible-a suitably difficult game so that the theories one comes up with are actually of some interest (unlike most of

their predecessors in the over-idealized philosophical tradition). In fact, these three constraints enable one to construct an argument with a conclusion surprisingly close to that of Professor Searle, about the importance of the brain in human thought. The constraints of android epistemology require one to construct a theory that is mechanistically realizable, but couched in very abstract terms-"software" terms, you might say. Such a theory describes what Searle calls a "purely formal system." (We might add, noting our third constraint, a purely formal *but* dynamic system, a system for which time is a critical parameter.) This leads then to a model of the mind or the knower composed in terms of strategies of formal operations and activities. That is in an ancient philosophical tradition, but now enriched with the new constraints. And it leads to a vision of what is important-or "essential"-about minds, vividly expressed by Maria Muldaur in a popular song:

It ain't the meat, it's the motion!

(I am indebted to Richard Sharvy for drawing my attention to this excellent use of Muldaur's song.) This might well be the motto of AI-or of what Searle calls "strong AI." Searle's own position is then succinctly captured:

It's the meat!

Here we have a sharp-edged difference of opinion. But the constraints of AI (or android epistemology) themselves provide the premises for an argument that comes close to resolving this disagreement with a dialectical compromise.

Probably many of you have read Edwin A. Abbott's amusing fantasy, *Flatland:* a Romance of Many *Dimensions/* which begins in a two-dimensional world--a plane inhabited by intelligent plane figures—triangles and other polygons. Someone--I cannot recall who--objected that this world was impossible (who ever thought otherwise!) because there could not be a two-dimensional intelligent being. In order to get sufficient connectivity in whatever played the role of the creatures' brains, there had to be three dimensions (so that wires could cross each other, in effect). John McCarthy points out to me that this is strictly false; John von Neumann proved years ago that a general automaton-a universal Turing machine-can be realized in two dimensions, but of course such an automaton trades speed for geometrical simplicity. One can vastly increase the speed of operation of a computer (or brain) by folding it back on itself and letting it communicate within itself in three dimensions. So it is no accident that our brains are three-dimensional. Moreover, one can now see with something approaching certainty that our brains need to be not just three-dimensional, but also organized for *parallel* processing. For many cognitive tasks-especially the pattern-detecting tasks of perception, and some memory-searching tasks-the "machine architecture" of a standard digital computer, a "von Neumann machine" that is organized to operate sequentially, doing just one thing **at**

a time, but doing each thing very fast, simply does not permit the right computations to be executed in the time available. So an intelligent being (a· being like us) must have a brain organized for very rich parallel processing-perhaps millions of channels wide. Still, it seems we could build such a device out of silicon chips with scarcely any major advance in technology.

But what if it turns out (as some think) that while the brain's ten billion (or so) neurons are the main *switching* elements of the mind, they do this ,by making essential use of small information-storing changes in their subcellular organic molecules. It may not be physically possible to mimic the information-handling powers of such collections of molecules with anything other than just such systems of molecules. (An allosteric enzyme molecule capable of considerable information processing weighs in at 10^{-17} grams, and if one tries simulating the behavior of groups of such molecules in other media, one soon creates a very large, very slow model).[2]

So it may very well turn out that the only way one can achieve the information-handling prowess of a human brain (in real time) is by using-a human brain! So it might turn out after all that the only way to have a mind like ours is to have a brain like ours, composed of the same organic materials, organized in roughly the same way. This leads to an apparent resolution of the disagreement between Searle and the proponents of "strong AI":

Probably, only this meat can give you that motion.

Has the disagreement now been dissolved? Searle, after all, has accused AI of ignoring the causal powers of the brain, and here is an argument, based on AI principles and constraints, showing how important the brain's actual causal powers are. But in fact, not only does this argument not resolve the disagreement, it throws into sharp relief the nature of Searle's curious and mystical view, and AI's reasons for resisting it.

As Professor Dreyfus noted in the panel discussion, Searle concedes that it is possible in principle to build a brain like device out of silicon chips (or other AI-approved hardware) that perfectly mimics the real time input-output behavior of a human brain! That is, you could throw a person's brain away, replace it with a suitably programmed computer (a "merely formal system" embodied in some inorganic hardware or other), and that person's body would go on behaving *exactly* as it would have gone on behaving had it kept its brain. The control powers of the brain are not the "causal powers" Searle makes so much of: in fact, the causal powers Searle admires are entirely independent of the information-receiving, information-processing, and controlling powers of the brain. A body controlled by a computer rather than a brain might seem to outside observers to be an intelligent person, with a mind like ours (in fact, it would pass the most demanding behavioral tests of intelligence we could devise), but there would not in fact be any mind there at all! Such a

computer, being a "purely formal system," would not "produce real intentionality," and hence there would exist not a glimmer of consciousness to associate with this animated (but--according to Searle--inanimate) body.

It has often been pointed out that Searle's view has a curious implication in the area of evolutionary biology. If, as he insists, a mindless ("purely formal") computer brain of this sort is possible, it presumably could have evolved by natural selection. Are we not lucky, then, that our ancestors did not happen to have one of those mindless brains instead of the brains we have! Since such brains would be input-output *equivalent* to ours, from the outside they would be indistinguishable; natural selection could find no leverage for selecting in favor of our conscious sort of brain, full of "intentionality," instead of selecting in favor of the zombie-computer sort of brain (full of the low-priced spread). If it had been our misfortune to have had mindless ancestors of that sort, we would now all be zombies!

Android epistemologists--the defenders of strong AI--declare that this imagined distinction between two sorts of otherwise behaviorally indistinguishable control system is illusory. The illusion is sustained by reflecting on what if anything "it would be like" to be (or have!) such a control system oneself. This insistence by Searle on what he calls the first-person point of view is, simply, a big mistake. It is the last gasp of Cartesian introspective certainty. As John McCarthy said in the panel discussion, the AI community regards its reliance on a third-person perspective as "virtuous."

Why should anyone be afraid of the third-person perspective? This idea of bringing the mind into the third-person, objective view of science strikes some people as an esthetically pleasing, promising idea: at last we are beginning to see, however dimly, a path uniting the last great outpost of mystery, the human mind, to the expanding dominion of science. Other people, however, see the idea as profoundly threatening and unsettling. As Professor Turkle has noted, many people exhibit quite strong emotional reactions when confronted with such suggestions.

I am reminded of the reaction that greeted Darwin's theory of evolution by natural selection. As we all know, Darwin's theory hit the world like a bolt of lightning. One of the curious facts about it was that its importance was widely misperceived by the public. People could feel in their bones that this new idea was somehow a terrible threat to their peace of mind, a nightmare dread come true, but in their anxiety, they fixed on the trivial implications of the new theory. Perhaps they were afraid to acknowledge its real force. People said, "He claims we are the cousins of apes!"-as if the presumed embarrassment of having hairy chattering ancestors in one's family tree were the worst blow to one’s self-image that the Darwinian theory could deliver. "He claims the story of creation in the Bible is false!" others charged, and this got closer to the heart of the matter, for it was ultimately the belief in God itself that was effectively undermined by Darwin's theory. Why? Because the most

compelling and sophisticated argument for the existence of God, the "argument from design," whose force could be appreciated by the most agnostic scientist, and which owed none of its appeal to faith or revelation or traditional dogmas-that best argument for the existence of God had suddenly lost its credentials. But even here, intelligent people missed, and perhaps chose to miss, the point.

Recall, for instance, the famous Scopes monkey trial, pitting William Jennings Bryan against Clarence Darrow. There was Bryan, the fundamentalist and Populist hero of the farm states, three times unsuccessful Democratic candidate for president, leading the prosecution of Scopes, who dared teach the theory of evolution to his high school students. The trial was one of the first historic events to receive the full attention of the modern media, with armies of reporters telegraphing hundreds of thousands of words about the trial to their newspapers each day. The nation was spellbound. But in all of that intense scrutiny, and in all of the oratory on both sides, the real challenge of Darwin's theory lay all but hidden.

Bryan put on quite a show. With vehement oratory, he laid down a few home truths that any simple man could understand. Any so-called scientist who could not see what any ordinary folk could see was just a fool. He ridiculed the opposition, taking care to find cheapened, oversimplified versions of the views he encouraged everyone to scoff at. We often diminish what we fear, hoping to turn what seems to be a bogeyman into a silly clown we will not have to fear anymore.

Today we see people picking up the jargon of computers and adapting it to popular culture, cheapening and diminishing it, and diverting their anxiety with hackneyed-and implausible-bogeymen. Would you like to marry a robot? Are you a (mere) Turing machine clanking away on a paper tape? Do you know someone who has been "deprogrammed" after being victimized by a religious cult? The fear is evident enough in many manifestations of the computer metaphor. Is it, like the earlier fear of the Darwinian metaphor, a misplaced anxiety? Is there in fact anything to the computer metaphor? Have we made any substantial progress in understanding the mind with the help of the concepts of computer science?

I will describe one very basic, very abstract contribution of computers to the understanding of the mind. There is a well-known purely conceptual problem which I call Hume's problem-not because David Hume solved it, but because he struggled mightily with it. You can find versions of it in recent writings by B. F. Skinner and Gilbert Ryle,[3,4] and in many others. There is a tremendous and plausible theoretical temptation to say that a mind is essentially something containing representations: memories, ideas, thoughts, sensations. Let us call those all mental representations. But a representation does not work, can play no role, unless there is a representation user or representation appreciator to manipulate it and to understand it. So a representation user has to be a mind. But if inside your mind you have representations, and a little representation user using those representations, what is inside that representation user's mind? More representations, with their own inner representation users? Do we

get an infinite regress of little men in the brain-"homunculi"-with littler men in their brains and so on? The whole idea of mental representation systems has thus seemed to many to be like the idea of a perpetual motion machine: a strictly impossible mechanism-a miraculous object.

Computer science has changed that. Not just AI but computer science in general, for a computer is, if nothing else, a mindless manipulator of representations. Well, maybe they are not representations, since they do not seem to need inner representation users with minds. Whatever they are, there is no infinite regress, for what is actual is possible, and there sit the computers, honest-to-goodness representation-using mechanisms. Inside these computers are something-or-others-"data structures" and other newfangled entities-that might well be called *self-understanding* representations. They are representations that understand themselves!

Now that might seem like a contradiction in terms, an obviously incoherent joining of concepts. But we should remember that there was a time when the concept of splitting the atom was equally self-contradictory. Atoms were by definition unsplittable. We have learned that there are such things as splittable atoms (though now perhaps we should agree that they are misnamed, precisely because they are splittable and do have parts-which their name, in Greek, denies). We have more recently learned that there are such things as self-understanding representations (though there may well be a better name for them). They perform the roles we traditionally assigned to the various dubious sorts of mental representations of earlier theories of the mind. And hence they break the back of the infinite regress argument that sets up Hume's problem.

All this shows, of course, is a certain possibility in principle. It does not establish how or why or in what ways a human mind might be like a computer. All it shows is that in principle we can have a theory of the mind that is mechanistic, finite, operates in real time, and is not a perpetual motion machine. There are still many fundamental conceptual problems (and an excellent survey of them was presented by Professor McCarthy), but progress is being made on them.

Should we, nevertheless, in the light of my earlier conclusion about the probable importance of the brain, abandon "pure" AI in favor of "bottom-up" brain research? No, for a reason that has been very clearly brought out by Professor Michael Rabin. No amount of parallel processing, no matter how many channels wide, and no amount of microminiaturization, even down to the level of organic molecules in individual nerve cells, is going to give one enough computational power to avoid a genuine exponential explosion of computation, a "combinatorial explosion" of information processing. There is only one way of avoiding that sort of frantic paralysis: clever software. Software-"purely formal systems”--is going to be required to make sense of the brain's operations in any case. We are going to have to understand how the strategies of representation and representation manipulation have achieved their

quite necessary economies, even if we are utilizing the full parallel resources of every neuron and every neuron part.

Some people still may feel the tug of fear at the prospect of taking this third-person perspective on the mind. Some may fear that we will somehow rob each other of what is special and wonderful about us. Some people may thus feel a strong desire to build a moat of some kind around some special part of them--their mind--and keep it forever inviolate and untouchable by science. All I can say to help to assuage that fear is that if you look at the previous great leaps forward in science you will see that far from diminishing our appreciation of the subtlety and wonder and complexity of the phenomena, they have increased it. Our knowledge of the genetic code and its operation, for instance, provides a far more spectacular vista on the nature of procreation than any of the un detailed vitalistic theories that preceded it. I for one have no doubt that if and when we ever do get a good third-person theory of the mind, it will only confirm our most optimistic sense of how extraordinarily complex and beautiful a human mind is.

REFERENCES

1. Abbott, E. A. 1962. Flatland: A Romance of Many Dimensions. Blackwell. Oxford. England.
2. Monod, J. 1971. Chance and Necessity: 69. Vintage/Random House. New York, N.Y.
3. Skinner, B. F. 1964. Behaviorism at fifty. In Behaviorism and Phenomenology. T. Wann, Ed.: 79-108. University of Chicago Press. Chicago, Ill.
4. Ryle, G. 1949. The Concept of Mind. Hutchinson. London, England .