DANIEL C. DENNETT

In Defense of AI

Before there was a field called cognitive science, I was involved in it when I was a graduate student at Oxford. At that time, I knew no science at all. I had had a purely humanistic education as an undergraduate. But I was very interested in the mind and in the philosophy of mind. I was completely frustrated by the work that was being done by philosophers, because they did not know anything about the brain, and they did not seem to be interested. So I decided that I had to start to learn about the brain to see what relevance it had. I became an autodidact neuroscientist, with the help of a few professors.

What I found out was that people who knew about brains did not have a lot to say about the mind, either. In those days, unlike today, it was very hard to come across much of anything in neuroscience that had the ambition of addressing any of the philosophical questions about mind.

Just about that time, I learned about Artificial Intelligence. This was in 1963–64. I got quite interested. There was a nice little book edited by Alan Anderson, *Minds and Machines*,[1] which was philosophical but raised some issues. I talked with Anderson who was in England that year. When I got my first job at the University of California at Irvine in 1965, there was a small Artificial Intelligence group there. One day one of them, Julian Feldman,[2] came into my office and threw a paper down on my desk, saying: "You're a philosopher; what do you make of this?" The paper he threw on my desk was Bert Dreyfus's first attack on Artificial Intelligence. It was his Rand memo, called "Alchemy and Artificial Intelligence."[3] I read it and thought it was wrong, that Dreyful made mistakes. He said he wanted me to write an answer, showing where the mistakes are.

Since it fit with my interests anyway, I wrote an article responding simultaneously to Dreyfus and to Allen Newell's view at that time. Newell had come to Irvine to give a talk, and I thought he was wrong, too, in a slightly different way. So I wrote a piece that responded to

Daniel C. Dennett was born in 1942. He earned his B.A. degree from Harvard University in 1963 and his D.Phil. from Oxford in 1965. From 1965 to 1971, he taught at the University of California at Irvine. He moved to Tufts University in Medford, Massachusetts, in 1971, where he has been since 1975. Since 1985, he has been Distinguished Professor of Arts and Sciences and Director of the Center for Cognitive Studies at Tufts University.

Dreyfus and Newell. It was actually my first publication, in 1968. It was an article published in the journal *Behavioral Science*, called "Machine Traces and Protocol Statements."[4]

That hooked me. I was very interested in the debate, and I was sure that Dreyfus was wrong. The people in Artificial Intelligence were glad to have a philosopher on their side. An interesting relationship began to develop, and it has continued over the years.

From where did you get your knowledge about Artificial Intelligence and computer science?

I got it out of interactions of this sort with people, growing gradually over the years, by talking with people and reading accessible literature. But I was not really computer-literate. In 1978–79, John McCarthy set up a group at the Center for Advanced Studies in Behavioral Science in Palo Alto on philosophy and Artificial Intelligence. That was a wonderful year. There were six of us there, all year long: John McCarthy, Zenon Pylyshyn, Patrick Hayes, Bob Moore, John Haugeland, and myself. Two philosophers and four Artificial Intelligence people.[5] In the course of the year, a lot of other people came around for short periods. I learned a tremendous amount; we talked a lot and debated a lot. I still did not come out of that year very computer-literate, but I made some considerable progress. In the years following, I developed that much further, even to the point where, a few years ago, I taught a computer science course here. I am not a serious programmer, but I know how to program, and I know some languages.

Was it difficult to discuss things with people in computer science, with your totally different background and training?

It has been difficult, but the surprising thing is that the real difficulty arises from the fact that although they are trained as computer scientists they use a lot of terms that philosophers use, and it takes a long time to discover that they don't mean the same things by them. Their terms are "false friends." You can have a conversation going on for quite a long time before you realize that you are talking past each other because you don't use the words in the same way. This was particularly apparent during the year at Stanford. It would be hard to find four more philosophical people in Artificial Intelligence than McCarthy, Pylyshyn, Hayes, and Moore. And it would be impossible to find two more Artificial Intelligence–minded philosophers than Haugeland and me. So you would think we could talk together very well.

You have the feeling that in this debate between Artificial Intelligence and its critics, you are on the side of Artificial Intelligence?

I am happy to identify myself as the defender of Artificial Intelligence, of strong Artificial Intelligence. I have been a critic of numerous errors I found in Artificial Intelligence, but on the overall issue, I think a very good case can be made for strong Artificial Intelligence, and a very good case can be made for the importance of details in the research.

You mentioned your criticism of Dreyfus. What are your main points in this issue?

When Dreyfus first started criticizing Artificial Intelligence, he did not know very much about the field, any more than I did. He had some good instincts about where the difficult problems for Artificial Intelligence lay, and he pointed them out. But he exaggerated. He claimed that things were impossible in principle, that these were obstacles that no sophistication in Artificial Intelligence could ever deal with. I thought his arguments were not plausible at all.

If you soften his claims and read him as saying that Artificial Intelligence has underestimated the difficulties of certain issues, that these are the outstandingly difficult problems for Artificial Intelligence, and that Artificial Intelligence would have to revolutionize some of its thinking in order to account for them—if that had been what he said, he would have been absolutely right.

I think there is still a great difference between your view and his. He is criticizing the whole rationalist tradition, whereas you come from this tradition and defend it.

Actually, I am not defending the particular line he is criticizing. Like most philosophers, Dreyfus tries to find something radical and absolute rather than a more moderate statement to defend. I think his current line about the bankruptcy of rationalism, if you like, is just overstated. Again, if you moderate it, I agree with it. I think that people in Artificial Intelligence, particularly people like John McCarthy, have preposterously overestimated what could be done with proof procedures, with explicit theorem-proving style inference. If that is what Dreyfus is saying, he is right.

But then, a lot of people in Artificial Intelligence are saying that, too, and have been saying it for years and years. Marvin Minsky, one of the founders of Artificial Intelligence, has always been a stern critic of that hyperrationalism of, say, McCarthy. Dreyfus does not really have a new argument if he is saying something moderate about rationalism. There is only a very limited and radical group of hyperrationalists who fall into this criticism. If, on the other hand, Dreyfus is making a claim that would have as its conclusion that all Artificial Intelligence, including connectionism and production systems, is bankrupt, he is wrong.

He is saying that our thinking is not only representational. There are things you cannot represent, like bodily movements, pattern recognition, and so on. I understand that he says the human mind is not merely a problem solver, as the physical symbol systems hypothesis would imply.

Everything in that statement is true except the claim at the end: that this is what the physical symbol systems hypothesis has to hold. Any sane version of the physical symbol systems hypothesis—that is not my way of thinking, but I defend Allen Newell on this ground—is quite prepared and able to grant that there are transitions in such a system that are not inferences. These transitions are changes from one state to another that are not captured by any informational term—that is, from the standpoint of the symbols, they are inexplicable. They are not the adding of a premise, they are not the adding of a hypothesis, they are not a revision of a representation—they are a change of state, but they cannot be captured in the language of representation. It is possible to ignore that possibility and think that you can have a theory of the mind that does not permit such changes. But this is a very strong and gratuitous claim, and there is no particular reason to hold this view.

In my own work, *The Intentional Stance*,[6] I give a number of examples of ways in which items, states, and events can have a determinate effect on a cognitive system without themselves being representations. In putting forward these examples, I was not arguing against any entrenched view of Artificial Intelligence.

So you think that one can combine the physical symbol systems hypothesis with the intuition that you cannot explain everything with a symbol?

Yes, I don't see why it could not be done. It is an artificial watershed that has been created as if you could not step across that line and be true to your view.

What about connectionism? It criticizes the physical symbol systems hypothesis and seeks alternative explanations for the human mind.

Indeed, connectionism is a perfect case of what we were talking about. It is the clearest vision we yet have of that mixture. If you look at the nodes in a connectionist network, some of them appear to be symbols. Some of them seem to have careers, suggesting that they are particular symbols. This is the symbol for "cat," this is the symbol for "dog." It seems likely to say that whenever cats are the topic, that symbol is active; otherwise it is not.

Nevertheless, if you make the mistake of a simple identification of these nodes—as cat symbol and dog symbol—this does not work. Because it turns out that you can disable this node, and the system can go

right on thinking about cats. Moreover, if you keep the cat and dog nodes going and disable some of the other nodes that seem to be just noisy, the system will not work. The competence of the whole system depends on the cooperation of all its elements, some of which are very much like symbols. There is no neat way of demarcating the events that are symbolic from the events that are not. At the same time, one can recognize that some of the things that happen to those symbols cannot be correctly and adequately described or predicted at the symbol level.

Even in this case, you would not say that there is a watershed between connectionism and the physical symbol systems hypothesis?

There is a big difference in theoretical outlook, but it is important to recognize that both outlooks count as Artificial Intelligence. They count as strong Artificial Intelligence. In John Searle's terms, connectionism is just as much an instance of strong Artificial Intelligence as McCarthy's or Schank's[7] or Newell's work. The last time I talked with him he was very clear about this—that you could be a strong Artificial Intelligence connectionist. He was just as skeptical about it as of any other kind of Artificial Intelligence.

Both Searle and Dreyfus generally think that it could be a more promising direction.

Sure, it is that. There is a bit of a dilemma for both of them. If they want to say: "Hurrah for connectionism, this is what we meant, this is what the alternative was supposed to be," then they are embarrassed to admit that it is strong Artificial Intelligence, just a different brand. Then their criticism was directed not at Artificial Intelligence in general but at a particular brand of Artificial Intelligence. On the other hand, if they say: "You are right, it is Artificial Intelligence, and Artificial Intelligence is impossible," then how can it be such a promising avenue? Searle could say that it is a promising avenue for weak Artificial Intelligence, but then we have to look and see whether he has any independent arguments against strong Artificial Intelligence that don't depend on his focus on the brand of Artificial Intelligence that is not connectionist.

Let's follow Searle's criticism of strong Artificial Intelligence. He has refuted strong Artificial Intelligence with his Chinese Room Argument. What do you think about that?

I think that is completely wrong. First of all, the Chinese Room is not an argument, it is a thought experiment. When I first pointed this out, Searle agreed that it is not an argument but a parable, a story. So it's neither sound nor an argument; it is merely a demonstration. I have pointed out in different places that it is misleading and an abuse of the

thought experiment genre. Searle's most recent reaction to all that is to make explicit what the argument is supposed to be. I have written a piece called "Fast Thinking," which is in my book *The Intentional Stance*. It examines that argument and shows that it is fallacious.

Briefly, Searle claims that a program is "just syntax," and because you can't derive semantics from mere syntax, strong Artificial Intelligence is impossible. But claiming that a program is just syntax is equivocal. In the sense that it is true, it is too strong for the use he makes of it (in the sense in which a program is just syntax, you can't even get word processing or weather forecasting out of it—and computers manifestly can perform those tasks). Otherwise, the premise is false; but beyond that, the claim that you can't derive semantics from syntax obscures a possibility that Artificial Intelligence has exploited all along: that you can approximate the performance of the ideal semantic engine (as I call it) by a device that is "just syntax."

One of the things that has fascinated me about the debate is that people are much happier with Searle's conclusion than with his path to the conclusion. For years I made the mistake in talking to people about the Chinese Room—it has been around for ten years now—of carefully showing what the fallacy is in the argument, in the sense that I said, "If this is an argument, it has to be *this* argument, which is fallacious, or it has to be *this* argument, which is fallacious," and so on. I would go very carefully through his arguments, but I think people were not interested. They did not care about the details of the argument, they just loved the conclusion. Finally I realized that the way to respond to Searle that met people's beliefs effectively was to look at the conclusion and ask what it actually is, to make sure that we understand that wonderful conclusion. When we do this, we find that it too is actually ambiguous.

There is another statement that is, at first appearance, very close to Searle's conclusion. This statement is no nonsense and might even be true. It is an interesting if not particularly likely empirical hypothesis: the only way to achieve the speed of computation required to re-create artificial intelligence is by using organic computation. The people may be thinking that this is what he is arguing for, but he is not. That is in my chapter "Fast Thinking."

I read your and Hofstadter's comment in The Mind's I.[8] *I think Searle would classify that reply as the "systems reply," because you argue that nobody claims that the microprocessor is intelligent but that the whole system, the room, behaves in an intelligent way. Searle modified the argument in response to this. He says: let's assume that all the system is in my head and that I learn everything by heart; then I would also behave intelligently. He says that the systems reply does not work because it is*

only one step further. He finds it strange that you would agree with him that the human being manipulating the symbols without knowing what they mean does not have intentionality but that the whole system—paper and pencil and all—does have intentionality.

This is not recent. If you go back to *Behavioral and Brain Sciences*,[9] you will find my commentary in there, which, at the time I wrote it, was already old. I had already been through that with Searle many times. In the little piece "The Milk of Human Intentionality" I point out that Searle does not address the systems reply and the robot reply correctly. He suggests that if he incorporates the whole system in himself, the argument still goes through. But he never actually demonstrates that by retelling the story in any detail to see if the argument goes through. I suggest in that piece that if you try it, your intuitions at least waver about whether or not there is anything there that understands Chinese or not.

We have got Searle, a Chinese-speaking robot, who has incorporated the whole system in himself. Suppose I encounter him in the street and say: "Hello, John, how are you?" What does your imagination tell you that happens? First of all, there seem to be two different systems there. One of them does not understand English—it is the Chinese robot. That system should reply to me something like "I am sorry, I don't speak any English!" We can tell the story in different ways. If we tell the story in a way in which he is so engrossed in manipulating his symbols that he is completely oblivious to the external world, then of course I would be certainly astonished to discover that my friend John Searle has disappeared and has been replaced by a Chinese-speaking, Chinese-understanding person. The sense that *somebody* understands Chinese is overpowering—and who else but Searle? We might say: it looks as if the agent Searle has been engulfed by a larger agent, Searle II, who understands Chinese. But that was the systems reply all along.

If Searle had actually gone to the trouble of looking at that version of his thought experiment, it would not have been as obvious to people anyhow.

Sure, you could ask the robot questions, observe its behavior, and ascribe it intentionality, as you wrote in your paper. But to ascribe intentionality does not mean that it really has intentionality.

That is an interesting claim. I have argued, of course, that that, in the limit, is all there is. There is not any original, intrinsic intentionality. The intentionality that gets ascribed to complex intentional systems is all there is.[10] It is an illusion that there is something more intrinsic or real. This is not a radical thesis, but a very straightforward implication of standard biological thinking. We are mechanisms, mechanisms with very elaborate purposes. Ultimately the raison d'être looked like a deci-

sive refutation of any representational theory for a long time, until we had computers.

When computers came along, we began to realize that there could be systems that did not have an infinite regress of homunculi but had a finite regress of homunculi. The homunculi were stupider and stupider and stupider, so finally you can discharge the homunculus, you can break it down into parts that are like homunculi in some regards, but replaceable by machines. That was one of the most liberating conceptual contributions of computers. It showed us how we could break the regress.

Notice that we have many examples of the regress broken, but we could argue, if there were any point to it, about how to describe the result. Should we say that computers don't really represent anything because there is nothing in them if you look at them? Or should we say that they are representations but they don't need complete minds inside to appreciate them? They can get by with systems that are only a little bit like homunculi. In any case, the regress is broken there.

Infinite regress seems to arise in other places. Let's take commonsense. If you think that commonsense is composed of a bunch of rules that are being followed, then you need a commonsense to know which rule to follow when. So you have to have metarules to tell you at which rules to look. But then you will have to have commonsense about how to use the metarules, and so on. If we try to do it with rules forever, we have an infinite regress of rules.

But we should learn from the first regress. This is not an infinite regress. It is simply an empirical question: how many layers of rules do you have to have before you get down to dispositions that can be replaced by a machine? This is an open question. It is strongly analogous to the question of how many layers of language there are between a particular computer application and the machine code. That is an empirical question, too: it may be hard-wired, it may be a virtual machine, it may be a virtual machine running on a virtual machine running on a virtual machine. . . . If it is the latter, there are several layers of directions, of rules, that the whole system is depending on. And indeed, it looks to see what it should do by consulting the next rule on the list. If you want to know how it does that, you may find some other rules. Finally you hit the bottom, and there are no more rules. You have reached the microcode, as they call it; now you are in the hardware.

So your view is, in general, optimistic?

In general, I think that infinite regress arguments are signs that point in the direction of their own solution. If you think about them, you find that it cannot be an infinite regress. So you start tinkering with the assumptions, and then you see how a finite regress is probably the answer.

What do you think are the major problems that cognitive science has to overcome at the moment?

We are just beginning to develop biologically plausible models of consciousness, and they suggest that the computational architecture of a brain capable of sustaining conscious thought is vastly more subtle and ingenious than the architectures we have developed so far. In particular, the standard insulation of function in engineer-designed architectures (each element has one task to perform) is almost certainly a feature that systematically prevents us from exploring the right part of design space. The brain is not magical, but its powers seem magical because they emerge from a tangle of multipurpose, semiautonomous, partially competitive elements. Working out the design principles of such systems is a major task facing Artificial Intelligence and computational neuroscience, but tantalizing insights are arising. Some of these are explored in my new book, *Consciousness Explained.*[11]

Recommended Works by Daniel C. Dennett

Brainstorms: Philosophical Essays on Mind and Psychology. Cambridge, Mass.: MIT Press, 1978.

Consciousness Explained. London: Penguin, 1992.

The Intentional Stance. Cambridge, Mass.: MIT Press, 1987.

(Ed., with D. Hofstadter.) *The Mind's I: Fantasies and Reflections on Self and Soul.* New York: Basic Books, 1981.

"Why the Law of Effect Will Not Go Away," "True Believers: The Intentional Strategy and Why it Works," "Making Sense of Ourselves," and "Quining Qualia." All in *Mind and Cognition: A Reader*, ed. W. G. Lycan. Oxford: Blackwell, 1991.

Notes

1. *Minds and Machines*, ed. A. R. Anderson (Englewood Cliffs, N.J.: Prentice-Hall, 1964).

2. At that time best known for the book co-written with E. A. Feigenbaum, *Computers and Thought* (New York: McGraw-Hill, 1963).

3. H. Dreyfus, "Alchemy and Artificial Intelligence" (Rand Corporation, Santa Monica, Calif., 1965, memo).

4. D. C. Dennett, "Machine Traces and Protocol Statements," *Behavioral Science* 13 (1968): 155–61.

5. John Haugeland is the other philosopher. McCarthy was founder and first director of the Artificial Intelligence labs at MIT and Stanford University. See, e.g., J. McCarthy, "Programs with Common Sense" in *Semantic Information Processing*, ed. M. Minsky (Cambridge, Mass.: MIT Press, 1968). pp. 403–18; and Z. W. Pylyshyn, "What the Mind's Eye Tells the Mind's Brain," *Psychological Bulletin* 8 (1973): 1–14. A major, more recent work by Pylyshyn is *Computation and Cognition: Toward a Foundation for Cogni-*

tive Science (Cambridge, Mass.: MIT Press, 1984); P. Hayes, "The Naive Physics Manifesto," in *Expert Systems in the Electronic Age*, ed. D. Michie (Edinburgh: Edinburgh University Press, 1969), pp. 463–502; "The Second Naive Physics Manifesto," in *Formal Theories of the Commonsense World*, ed. J. R. Hobbs and R. C. Moore (Norwood, N.J.: Ablex, 1985), pp. 1–36. R. C. Moore, "The Role of Logic in Knowledge Representation and Commonsense Reasoning" in *Readings in Knowledge Representation*, ed. R. J. Brachman and H. J. Levesque (San Mateo, Calif.: Morgan Kaufmann, 1985), pp. 336–41.

6. D. C. Dennett, *The Intentional Stance* (Cambridge, Mass.: MIT Press, 1987).

7. R. C. Schank, *Conceptual Information Processing* (Amsterdam: North-Holland, 1975); *Dynamic Memory: A Theory of Learning in Computers and People* (Cambridge: Cambridge University Press, 1982); R. C. Schank and R. Abelson, *Scripts, Plans, Goals, and Understanding* (Hillsdale, N.J.: Lawrence Erlbaum, 1977).

8. D. R. Hofstadter, "Reflections on John R. Searle: Minds, Brains, and Programs" in *The Mind's I* by D. R. Hofstadter and D. C. Dennett (New York: Basic Books, 1981).

9. The journal *Behavioral and Brain Sciences* 3 (1980), where Searle's article containing the Chinese Room Argument "Minds, Brains, and Programs" was first published, together with a series of replies.

10. See the chapter "Intentional Systems" in D. C. Dennett, *Brainstorms* (Cambridge, Mass.: MIT Press, 1978).

11. D. C. Dennett, *Consciousness Explained* (Boston: Little Brown, 1991).