The New York Review of Books

VOLUME 29, NUMBER 11 · JUNE 24, 1982

The Myth of the Computer: An Exchange

By Daniel C. Dennett, Reply by John R. Searle

In response to The Myth of the Computer (APRIL 29, 1982)

To the Editors:

In *The Mind's I*, Douglas Hofstadter and I reprint (correctly) John Searle's much-discussed article, "Minds, Brains, and Programs," and follow it with a "Reflection" that is meant to refute his position, as he notes in his review [*NYR*, April 29]. Searle charges that in that Reflection we "fabricate a direct quotation" which moreover "runs dead opposite" to what he in fact says. The Pocket *OED* says "fabricate" means "invent (lie, etc.); forge (document)" so Searle is suggesting (at some length) that this is a deliberate misquotation—a very serious charge which we categorically deny.

Here are the facts. We *do* misquote him in the Reflection, alas; we have him saying "a few slips of paper" where he in fact says "bits of paper." This misquotation was enitrely inadvertent; we apologize to him for it; we have arranged for the error to be corrected in any future printings of the book.

Now, does the error make a difference worth mentioning? Searle claims it does. He claims that the misquotation is "the basis" of our argument, which could not proceed without it, since it "runs dead opposite" to his meaning. We do, as he says, repeat the error five times! (In effect, we got off on the wrong foot and then quoted our own error four times.) But so little does our case depend on the *mis*quotation, that once it is corrected no further revision—not so much as a word or comma—of our Reflection is called for or contemplated.

How could Searle think "a few slips of paper" differs so dramatically from "bits of paper"? We had better look at the context from which we have (mis)taken the fatal phrase. Here is what Searle says, as printed correctly on p. 359 of *The Mind's I*:

> The idea is that while a person doesn't understand Chinese, somehow the *conjunction* of that person and bits of paper might understand Chinese. It is not easy for me to imagine how someone who was not in the grip of an ideology would find that idea at all plausible.

Here Searle is ridiculing what he calls "the systems reply" to his view, and as he admits, he has a hard time taking it seriously. That is one of the points we were trying to make. He also says, in his review: "The mental gymnastics that partisans of strong AI have performed in their attempts to refute this rather simple argument [his "Chinese Room" thought experiment] are truly extraordinary." Here we have the spectacle of an eminent philosopher going around the country trotting out a "rather simple argument" and then marveling at the obtuseness of his audiences, who keep trying to show him what's wrong with it. He apparently cannot bring himself to contemplate the possibility that *he* might be missing a point or two, or underestimating the opposition. As he notes in his review, no less than twenty-seven rather eminent people responded to his article when it first appeared in *Behavioral and Brain Sciences*, but since he repeats its claims almost verbatim in the review, it seems that the only lesson he has learned from that response was that there are several dozen fools in the world. (Several dualists, including Sir John Eccles, the Nobel laureate neurophysiologist, sided with Searle.)

We claim that he has frankly misunderstood the systems reply, and that his remark about "bits of paper" betrays this—and has "blinded him to the realities of the situation." Sometimes it even seems as if he deliberately misrepresents the systems reply, as when he says in his review: "Adherents of this view believe, to my constant amazement, that though man fails to understand, the *room* understands Chinese." Searle's amazement stops just short of inspiring any doubt in his mind about the fidelity of his interpretation, but perhaps this is to be explained by a certain exegetical carelessness rather than willful caricature.

What is the heart of the systems reply? It is a distinction of levels that is not at all mysterious, or new, though Searle's diminutive "bits of paper" acts to minimize (or obfuscate) the point. "The *conjuction* of a person and bits of paper" doesn't *sound* like a very different system from a person alone, does it? How about "the conjunction of a person and the Library of Congress with its attendant staff"? Does that sound like a supersystem that just *might* have some interesting powers or properties lacked by any of its proper parts or subsystems? The latter comparison should suit Searle just fine, if (as he now claims) he meant his "bits of paper" to carry no diminutive implications. And it is fairer, since Searle is supposedly talking about an imagined super-program that passes any and all Turing tests, a program many orders of magnitude grander than anything yet written.

Searle, in a letter to me (which he has kindly permitted me to quote), says:

> In any case you and Hofstadter still miss the point. No matter how big the program, the conjunction of man and bits of paper is no different from man alone. All of the bits of paper in the world add nothing to the neurophysiological powers of the man's brain. The whole point of reminding the reader that these are just "bits of paper" is that they are not in any way an addition to the specific neurophysiological powers of the man's brain.

Here Searle manifestly misunderstands the systems reply. No one claims the supersystem gives the subsystem by itself special new powers or properties. Rather, we (and many others) claim that the supersystem itself—the whole supersystem—has these powers. Searle's persistent deaf ear to this point puzzles me, particularly since it is really just a "category mistake" claim of the sort that was all the rage during Searle's graduate student days at Oxford. In his reply to my earlier commentary on his paper (in *Behavioral and Brain Sciences*) he objects to my rather Oxonian claim that *I* understand English—my brain doesn't—with the retort: "I find his claim as implausible as insisting, 'I digest pizza; my stomach and my disgestive tract don't.' " How important a single word can be! The verb "digest" is nicely chosen, for note how radically the image shifts if we switch to "eat" or "enjoy." Does Searle find it quite all right to say that his stomach eats pizza? Can his mouth eat pizza? Which proper part of him could be said to enjoy the pizza? Levels do make a difference. Anyone who hunts for a pizza-enjoying subsystem in a human being is on a fool's errand, and anyone who denies that a supersystem understands Chinese on the grounds that none of its subsystems do is making the same error moving in the other direction.

This error is hidden in the flurry (or is it a mountain?) of bits of paper. Searle's original article abounds in misdirection of this sort. Is it deliberate or inadvertent? Searle objects to our giving him the benefit of the doubt and calling his phrase "casual" and "offhand." Would he prefer us to call it deliberate misdirection? In my earlier commentary in *Behavioral and Brain Sciences* I described his article as "sophistry," but Hofstadter and I took a more charitable line in our volume. We, unlike Searle, do not pretend to be able to divine intention in the slips of our opponents.

We are sorry we slipped over "a few slips," but if Searle actually thinks this was a deliberate "fabrication"—or that our case against his view depends on misquotation—he has deluded himself.

As for the rest of Searle's review, it contains much to which we object, but we have pre-refuted virtually all of it, point by point, in the book he was reviewing. Indeed, Searle's review is, with perhaps one novelty, simply a telescoped version of his article. Searle may think that "Say it again, faster, in the pages of *The New York Review*" is a sound tactic of persuasion, but we don't. So for the most part we are content to refer readers who want to figure out what is wrong with Searle's view to our book. The one somewhat new element in the review is the enlargement on his unusual idea that we ignore the "causal powers of the brain," and since one can easily misread Searle on this point, a little clarification is in order.

Searle stresses that a computer program, being "purely formal," has no causal powers of its own. True, but of course when a program is physically realized in some hardware, and attached by "transducers and effectors" to relevant portions of the rest of the world, that physically realized program can have lots of causal powers: such a program can control an oil refinery, make out payroll checks or—terrible to say—guide nuclear missiles to their targets. Let's call such causal powers *control powers*. Such powers are not simulated but real; the computer doesn't simulate controlling the refinery; it really does control the refinery. (The distinction between simulating and duplicating is not as unproblematic as Searle supposes, but we will give him the distinction here for the sake of argument.)

Now Searle has admitted (in conversation on several occasions) that in his view a computer program, physically realized on a silicon chip (or for that matter a beer-can contraption suitably sped up and hooked up) could in principle *duplicate*—not merely simulate—the control powers of the human brain. That is, such a computer program (somehow realized) could control a human body in all its activities. Would such a body have a mind? We on the outside would find its behavior indistinguishable from that of a normal human being, but whether or not it *really* had a mind would depend, Searle insists, on whether the hardware realization of the control program shared with the missing brain not only all its control powers (granted *ex hypothesi*) but also some *other* "causal powers" entirely undetectable by others in behavior, including the behaviors of introspective speech, emotional reaction, and so forth.

What powers could these be? Where would the physical *effects* of these neurophysiological powers show up? Searle answers that they would show up in the individual subject's consciousness of his own intentionality. But would these be physical effects? If so, they must be detectable (in principle) by outsiders. Would they register on the instruments of neuroscientists (if not "behaviorists")? Searle does not say, but since he insists that the effects are introspectible (only?) it is tempting to conclude that the effects are presumed to be non-physical, and that Searle is some sort of dualist. He adamantly denies it; he insists the causal powers he is discussing are physical, so they must have physical, publicly observable effects. Where, if not in the subject's behavior? Just in the brain? What would these effects *do*?

These are mysterious causal powers indeed, despite their scientific-sounding name. We frankly disbelieve in them—which is the extent of our "behaviorism." Surely we all agree that anything that has all the relevant causal powers of food—it saves one from starving, sustains growth and repair, tastes good, etc.—*is* food. And anything that has all the causal powers of oxygen is oxygen. We think that you could in principle give a body an artificial brain by giving it something that duplicated *all* the brain's *control* powers. And any creature so equipped would "have a mind" in the only sense that makes any sense: it would have a well-functioning (prosthetic) brain. Now perhaps we are wrong; perhaps there are some other causal powers that matter. Searle thinks so; he thinks organic brains "produce intentionality." It sometimes seems as if he thinks intentionality is some marvelous fluid secreted by the brain—but we shrink from imputing such a silly view to him, and await his further clarification of his position.

Searle paints us as taken in by the "mythology" of computers. We see ourselves as demythologizers, and Searle as the victim of several superannuated myths, but perhaps we have misinterpreted his view.

Daniel C. Dennett

Tufts University

Medford, Massachusetts

John R Searle replies:

I am glad Dennett acknowledges that he and his co-editor misquoted me five times, but I do not agree that the misquotations make no difference to their argument. On the contrary their version of the "systems reply" makes essential use of the presumed size and complexity of a computer program for understanding Chinese. I really would have been "blinded" to the "realities of the situation" if I had thought that the program consisted in "a *few* slips of paper," but in fact the statement of the systems reply given by me (p. 358) makes it clear that the program would occupy *a very large number* of bits of paper, which is dead opposite to the view they attribute to me. As Dennett says, "How important a single world can be!"

I am also glad that Dennett does not contest the interpretation that I have given of their views for he has now accepted what I presented as a *reductio ad absurdum* of their position. On their view a system of beer cans, appropriately programmed and with the right inputs and outputs, would have exactly the same mental states and processes that the human brain enables human beings to have. And the point is not just that the beer-can system would *simulate* having mental states or that it *might for all we know* have mental states but that it *must have* the same mental states, it must, e.g., feel thirsty, worry about itemized deductions, want to go to the bathroom, or think Proust is a better writer than Balzac.

How did Dennett and Hofstadter ever get into such an implausible position? It is a direct logical consequence of their acceptance of the three theses I have called "strong AI": the mind is just a computer program, the specific neurophysiology of the brain is irrelevant to the mind, and any behavior that satisfies the Turing test is conclusive proof of the presence of mental states. I argued that the first and third of the theses are demonstrably false, and once you abandon them there is not much point to the second. The Chinese room argument shows that an agent could have any formal program you like and could pass the Turing test for understanding Chinese and still not understand a word of Chinese. And the reason for this is that *the person in the Chinese room has syntax but no semantics*, i.e., he has Chinese symbols and rules for manipulating them but no way of attaching any meaning to the Chinese symbols.

Now Dennett believes that the systems reply refutes my argument and that I "manifestly misunderstand" the systems reply. This is clearly the crux of his letter so let us turn our full attention to the systems reply and see what it does and does not establish. The reply claims that though the man in the room does not understand Chinese the whole "supersystem" of which he is a part really does understand Chinese (the idea that the *room* then understands Chinese, by the way, was stated to me by one of the early inventors of the systems reply). Dennett thinks my only reply to this is to deny the distinction between supersystem and subsystem and to deny "that a supersystem understands Chinese on the grounds that none of its subsystems do." But that is not and has never been my argument, neither in my published writings nor in my letter to him. My objection to the systems reply is that though there is a quite valid distinction between the level of the supersystem and the level of the subsystem, it is irrelevant to the issue because neither level has any way of attaching any meanings to the Chinese symbols:

> The obvious objection to this [the systems reply] is that the system has no way of attaching meaning to the uninterpreted Chinese symbols, any more than the men did in the first place. The system, like man, has 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.

Dennett pretends that the issue between us is over the possibility of the supersystem having properties which are not properties of any of its subsystems. But that is not the issue. The issue, to repeat, is over how the supersystem—even a supersystem as big as the Library of Congress—can attach any meaning to any of the symbols. And neither Dennet nor Hofstadter nor any other partisan of the systems reply has even begun to show how it could.[*] The normal speaker of Chinese has brain capacities that enable him to attach meanings to Chinese symbols; and the point of the passage Dennett quotes from my letter to him was not to answer the systems reply but to point out that these specific neurophysiological capacities of the brain are not duplicated by the supersystem in the example, because the bits of paper add nothing to the specific neurophysiological powers of the brain of the non-Chinese-speaking man.

Dennett fears that my denial of behaviorism must lead me into some sort of dualism and he thinks that the causal powers I attribute to the brain are "mysterious." But there is no dualism and even less mystery. As far as we know, here is how it works. Mental states and processes, e.g., feeling thirsty or having a visual experience, are both *caused by* and *realized in* the neurophysiology of the brain. They are "caused by" in the unmysterious sense that conscious feelings of thirst and conscious visual experiences are the results of neuronal processes. They are "realized in" in the sense that they are right there in the brain, real conscious features of the brain, though of course at a higher level of description than that of the individual neuron. They are not some extra juice secreted by the neurons. In nature it is very common to find higher-level phenomena both caused by and realized in lower-level phenomena. The liquidity of the water I drink and the solidity of the table I am working on are both caused by the molecular behavior of and realized in the molecular structure of the systems in question. Neither solidity nor liquidity nor intentionality are "juices" secreted by microstructures. They are, rather, real features of those systems at a higher level of description than that of the microstructure. As Dennett says, "levels do make a difference."

But, argues Dennett, isn't it an absurd consequence of my view that two systems might have the same control powers, might exhibit the same behavior, and yet one have mental states and the other one not? There is nothing absurd about this consequence. A mechanical robot rigged up to a system of beer cans could, in principle, simulate human behavior exactly and still not have any mental states. This is no more mysterious than the fact that both a steam locomotive and an electric locomotive can pull a train at the same speed over the same distance while operating on quite different internal principles. In general two systems can produce the same external effects while working on quite different internal principles, and Dennett's behaviorism prevents him from seeing this because it leads him to concentrate solely on the external effects and on the "control powers."

Dennett assures us that in their book he and his co-editor have "pre-refuted virtually all" of the criticisms I make of them "point by point." That, I fear, is pure bluff. Except for the misquotation and the mistakes and misunderstandings I pointed out in my review they have added nothing to objections I have already answered. In particular they have no answer to the two most serious criticisms I make. Namely, it is a consequence of their view that any hunk of junk whatever would literally have to have mental states in the same sense that you and I do if only it instantiated the appropriate program with the right input and output; and they have absolutely no answer to the objection that their computer programs have no way of getting from syntax to semantics, they have no way of attaching mental content to formal features.

On my view it is just a plain (testable, empirical) fact about the world that it contains certain biological systems, specifically human and certain animal brains, that are capable of causing mental phenomena with intentional or semantic content. It is a trivial consequence of this fact that anything else that was capable of causing mental phenomena would have to have causal powers equivalent to these brains; and I present a separate argument to show that no formal computer program *by itself* would ever be sufficient to produce these causal powers. On my view it is an *objective* fact that the world contains *subjective* phenomena, and a *physical* fact that it contains *mental* phenomena. All of this is denied by Dennett, Hofstadter, and other partisans of strong AI.

Computers are useful, indeed increasingly indispensable, in psychology and biology as they are in other sciences. Perhpas this is a good place to express my enthusiasm for the prospects of weak AI, the use of the computer as a tool in the study of the mind. But what are we to make of the strong AI belief that the appropriately programmed computer literally has a mind, and its antibiological claim that the specific neurophysiology of the brain is irrelevant to the study of the mind? Notice that the views of strong AI are well financed and backed by prestigious teams of research workers. What should our response to these views be? I believe that strong (as distinct from weak) AI is simply play acting at science, and my aim both in my original articles and in this letter has been the relentless exposure of its preposterousness.

Notes

[*] Actually Dennett couldn't have shown how to get from the syntax to the semantics, from form to mental content, because his brand of behaviorism makes it impossible for him to accept the existence of semantics or mental

contents literally construed. He believes that nothing *literally* has any intrinsic intentional mental states, that when we say of someone that he has such mental states we are just adopting a certain "stance" toward him and his behavior, the "intentional stance."

Copyright © 1963-2007 NYREV, Inc. All rights reserved. Nothing in this publication may be reproduced without the permission of the publisher.