possible in principle to build machines that make plans and achieve goals. Searle has given us no evidence that this is not possible.

*by* **Arthur C. Danto**

*Department of Philosophy, Columbia University, New York, N.Y. 10027*

## The use and mention of terms and the simulation of linguistic understanding

In the ballet *Coppélia,* a dancer mimics a clockwork dancing doll simulating a dancer. The imitating movements, dancing twice removed, are predictably "mechanical," given the discrepancies of outward resemblance between clockwork dancers and real ones. These discrepancies may diminish to zero with the technological progress of clockwork, until a dancer mimicking a clockwork dancer simulating a dancer may present a spectacle of three indiscernible dancers engaged in a pas de trois. By behavioral criteria, nothing would enable us to identify which is the doll, and the lingering question of whether the clockwork doll is really dancing or only seeming to seems merely verbal – unless we adopt a criterion of meaning much favored by behaviorism that makes the question itself nonsensical.

The question of whether machines instantiate mental predicates has been cast in much the same terms since Turing, and by tacit appeal to outward indiscernibility the question of whether machines *understand* is either dissolved or trivialized. It is in part a protest against assimilating the meaning of mental predicates to mere behavioral criteria – an assimilation of which Abelson and Schank are clearly guilty, making them behaviorists despite themselves – that animates Searle's effort to mimic a clockwork thinker simulating understanding; to the degree that he instantiates the same program it does and *fails* to understand what is understood by those whom the machine is designed to simulate – even if the output of the three of them cannot be discriminated – then the machine itself fails to understand. The argumentation is picturesque, and may not be compelling for those resolved to define (such terms as) "understanding" by outward criteria. So I shall recast Searle's thesis in logical terms which must force his opponents either to concede machines do not understand or else, in order to maintain they *might* understand, to abandon the essentially behaviorist theory of meaning for mental predicates.

Consider, as does Searle, a language one does not understand but that one can in a limited sense be said to *read.* Thus I cannot read Greek with understanding, but I know the Greek letters and their associated phonetic values, and am able to pronounce Greek words. Milton's daughters were able to read aloud to their blind father from Greek, Latin, and Hebrew texts though they had no idea what they were saying. And they could, as can I, answer certain questions about Greek words, if only how many letters there are, what their names are, and how they sound when voiced. Briefly, in terms of the distinction logicians draw between the *use* and *mention* of a term, they knew, as I know, such properties of Greek words as may be identified by someone who is unable to use Greek words in Greek sentences. Let us designate these as M-properties, in contrast to U-properties, the latter being those properties one must know in order to use Greek (or any) words. The question then is whether a machine programmed to simulate understanding is restricted to M-properties, that is, whether the program is such that the machine cannot *use* the words it otherwise may be said to manipulate under M-rules and M-laws. If so, the machine exercises its powers over what we can recognize in the words of a language we do not understand, without, as it were, thinking *in* that language. There is some evidence that in fact the machine operates pretty much by pattern recognition, much in the manner of Milton's unhappy daughters.

Now I shall suppose it granted that we cannot define the U-properties of words exhaustively through their M-properties. If this is true, Schank's machines, restricted to M-properties, cannot think *in* the languages they *simulate* thinking in. One can ask whether it is possible for the machines to exhibit the output they do exhibit if all they have is M-competence. If not, then they must have some sort of U-competence. But the difficulty with putting the question thus is that there are two ways in which the output can be appreciated: as showing understanding or as only seeming to, and as such the structure of the problem is of a piece with the structure of the mind-body problem in the following respect. Whatever outward behavior, even of a human being, we would want to describe with a psychological (or mental) predicate – say that the *action* of raising an arm was performed – has a physical description that is true whether or not the psychological description is true – for example, that the arm went up. The physical description then underdetermines the distinction between bodily movements and actions, or between actions and bodily movements that exactly resemble them. So whatever outward behavior takes a (psychological) $\Psi$-predicate takes a (physical) $\Phi$-predicate that underdetermines whether the former is true or false of what the latter is true of. So we cannot infer from a $\Phi$-description whether or not a $\Psi$-description applies. To be sure, we can ruthlessly define $\Psi$-terms as $\Phi$-terms, in which case the inference is easy but trivial, but then we cannot any longer, as Schank and Abelson wish to do, *explain* outward behavior with such concepts as understanding. In any case, the distinction between M-properties and U-properties is exactly parallel: anything by way of output we would be prepared to describe in U-terms has an M-description true of it, which underdetermines whether the U-description is true or not.

So no pattern of outputs entails that language is being used, nor hence that the source of the output understands, inasmuch as it may have been cleverly designed to emit a pattern exhaustively describable in M-terms. The problem is perfectly Cartesian. We may worry about whether any of our fellows is an automaton. The question is whether the Schank machine (SAM) is so programmed that only M-properties apply to its output. Then, however closely (exactly) it simulates what someone with understanding would show in his behavior, *not one step* has been taken toward constructing a machine that understands. And Searle is really right. For while U-competence cannot be defined in M-terms, an M-specified simulation can be given of any U-performance, however protracted and intricate. The simulator will only show, not have, the properties of the U-performance. The performances may be indiscriminable, but one constitutes a use of language only if that which emits it in fact uses language. But it cannot be said to use language if its program, as it were, is written solely in M-terms.

The principles on the basis of which a user of language structures a story or text are so different from the principles on the basis of which one could predict, from certain M-properties, what further M-properties to expect, that even if the outputs are indiscernible, the *principles* must be discernible. And to just the degree that they deviate does a program employing the latter sorts of principles fail to simulate the principles employed in understanding stories or texts. The degree of deviation determines the degree to which the strong claims of AI are false. This is all the more the case if the M-principles are not to be augmented with U-principles.

Any of us can predict what sounds a person may make when he answers certain questions that he understands, but that is because we understand where he is going. If we had to develop the ability to predict sounds only on the basis of other sounds, we might attain an astounding congruence with what our performance would have been if we knew what was going on. Even if no one could tell we didn't, understanding would be nil. On the other hand, the question remains as to whether the Schank machine *uses* words. If it does, Searle has failed as a simulator of something that does not simulate but genuinely possesses understanding. If he is right, there is a pretty consequence. M-properties yield, as it were, pictures of words: and machines, if they encode propositions, do so pictorially.

*by* **Daniel Dennett**

*Center for Advanced Study in the Behavioral Sciences, Stanford, Calif. 94305*

## The milk of human intentionality

I want to distinguish Searle's arguments, which I consider sophistry, from his positive view, which raises a useful challenge to AI, if only becuase it should induce a more thoughtful formulation of AI's foundations. First, I must support the charge of sophistry by diagnosing, briefly, the tricks with mirrors that give his case a certain spurious plausibility. Then I will comment briefly on his positive view.

Searle's form of argument is a familiar one to philosophers: he has

constructed what one might call an *intuition pump*, a device for provoking a family of intuitions by producing variations on a basic thought experiment. An intuition pump is not, typically, an engine of discovery, but a persuader or pedagogical tool – a way of getting people to see things *your* way once you've seen the truth, as Searle thinks he has. I would be the last to disparage the use of intuition pumps – I love to use them myself – but they can be abused. In this instance I think Searle relies almost entirely on ill-gotton gains: favorable intuitions generated by misleadingly presented thought experiments.

Searle begins with a Schank-style AI task, where both the input and output are linguistic objects, sentences of Chinese. In one regard, perhaps, this is fair play, since Schank and others have certainly allowed enthusiastic claims of understanding for such programs to pass their lips, or go uncorrected; but from another point of view it is a cheap shot, since it has long been a familiar theme *within AI circles* that such programs – I call them *bedridden* programs since their only modes of perception and action are linguistic – tackle at best a severe truncation of the interesting task of modeling real understanding. Such programs exhibit no "language-entry" and "language-exit" transitions, to use Wilfrid Sellars's terms, and have no capacity for non linguistic perception or bodily action. The shortcomings of such models have been widely recognized for years in AI; for instance, the recognition was implicit in Winograd's decision to give SHRDLU something to do in order to have something to talk about. "A computer whose only input and output was verbal would always be blind to the meaning of what was written" (Dennett 1969, p. 182). The idea has been around for a long time. So, many if not all supporters of strong AI would simply agree with Searle that in his initial version of the Chinese room, no one and nothing could be said to understand Chinese, except perhaps in some very strained, elliptical, and attenuated sense. Hence what Searle calls "the robot reply (Yale)" is no surprise, though its coming from Yale suggests that even Schank and his school are now attuned to this point.

Searle's response to the robot reply is to revise his thought experiment, claiming it will make no difference. Let our hero in the Chinese room also (unbeknownst to him) control the nonlinguistic actions of, and receive the perceptual informings of, a robot. Still (Searle asks you to consult your intuitions at this point) no one and nothing will really understand Chinese. But Searle does not dwell on how vast a difference this modification makes to what we are being asked to imagine.

Nor does Searle stop to provide vivid detail when he again revises his thought experiment to meet the "systems reply." The systems reply suggests, entirely correctly in my opinion, that Searle has confused different levels of explanation (and attribution). *I* understand English; my brain doesn't – nor, more particularly, does the proper part of it (if such can be isolated) that operates to "process" incoming sentences and to execute my speech act intentions. Searle's portrayal and discussion of the systems reply is not sympathetic, but he is prepared to give ground in any case; his proposal is that we may again modify his Chinese room example, if we wish, to accommodate the objection. We are to imagine our hero in the Chinese room to "internalize all of these elements of the system" so that he "incorporates the entire system." Our hero is now no longer an uncomprehending *sub*-personal part of a supersystem to which understanding of Chinese might be properly attributed, since there is no part of the supersystem external to his skin. Still Searle insists (in another plea for our intuitional support) that no one – not our hero or any *other* person he may in some metaphysical sense now be a part of – can be said to understand Chinese.

But will our intuitions support Searle when we imagine this case in detail? Putting both modifications together, we are to imagine our hero controlling both the linguisitc and nonlinguistic behavior of a robot who is – himslef! When the Chinese words for "Hands up! This is a stickup!" are intoned directly in his ear, he will uncomprehendingly (and at breathtaking speed) hand simulate the program, which leads him to do things (*what* things – is he to order himself in Chinese to stimulate his own motor neurons and then obey the order?) that lead to his handing over *his own* wallet while begging for mercy, in Chinese, with his own lips. Now is it at all obvious that, imagined this way, no one in the situation understands Chinese? In point of fact, Searle has simply not told us how he intends us to imagine this case, which we are licensed to do by his two modifications. Are we to suppose that if the words had been in English, our hero would have responded (appropriately) in his native English? Or is he so engrossed in his massive homuncular task that he responds with the (simulated) incomprehension that would be the program-driven response to this bit of incomprehensible ("to the robot") input? If the latter, our hero has taken leave of his English-speaking friends for good, drowned in the engine room of a Chinese-speaking "person" inhabiting his body. If the former, the situation is drastically in need of further description by Searle, for just what he is imagining is far from clear. There are several radically different alternatives – all so outlandishly unrealizable as to caution us not to trust our gut reactions about them in any case. When we imagine our hero "incorporating the entire system" are we to imagine that he pushes buttons with his fingers in order to get his own arms to move? Surely not, since all the buttons are now internal. Are we to imagine that when he responds to the Chinese for "pass the salt, please" by getting his hand to grasp the salt and move it in a certain direction, he doesn't *notice* that this is what he is doing? In short, could anyone who became accomplished in this imagined exercise fail to become fluent in Chinese in the process? Perhaps, but it all depends on details of this, the only crucial thought experiment in Searle's kit, that Searle does not provide.

Searle tells us that when he first presented versions of this paper to AI audiences, objections were raised that he was prepared to meet, in part, by modifying his thought experiment. Why then did he not present us, his subsequent audience, with the modified thought experiment in the first place, instead of first leading us on a tour of red herrings? Could it be because it is impossible to tell the doubly modified story in anything approaching a cogent and detailed manner without provoking the *unwanted* intuitions? Told in detail, the doubly modified story suggests either that there are two people, one of whom understands Chinese, inhabiting one body, or that one English-speaking person has, in effect, been engulfed within another person, a person who understands Chinese (among *many* other things).

These and other similar considerations convince me that we may turn our backs on the Chinese room at least until a better version is deployed. In its current state of disrepair I can get it to pump my contrary intuitions at least as plentifully as Searle's. What, though, of his positive view? In the conclusion of his paper, Searle observes: "No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle." I don't think this is just a curious illustration of Searle's vision; I think it vividly expresses the feature that most radically distinguishes his view from the prevailing winds of doctrine. For Searle, intentionality is rather like a wonderful substance secreted by the brain the way the pancreas secretes insulin. Brains *produce intentionality*, he says, whereas other objects, such as computer programs, do not, even if they happen to be designed to mimic the input-output behavior of (some) brain. There is, then, a major disagreement about what the *product* of the brain is. Most people in AI (and most functionalists in the philosophy of mind) would say that its product is something like *control:* what a brain is *for* is for governing the right, appropriate, intelligent input-output relations, where these are deemed to be, in the end, relations between sensory inputs and behavioral outputs of some sort. That looks to Searle like some sort of behaviorism, and he will have none of it. Passing the Turing test may be prima facie evidence that something has intentionality – really has a mind – but "as soon as we knew that the behavior was the result of a formal program, and that the actual causal properties of the physical substance were irrelevant we would abandon the assumption of intentionality."

So on Searle's view the "right" input-output relations are symptomatic but not conclusive or criterial evidence of intentionality; the proof of the pudding is in the presence of some (entirely unspecified) causal properties that are *internal* to the operation of the brain. This internality needs highlighting. When Searle speaks of causal properties one may

think at first that those causal properties crucial for intentionality are those that link the activities of the system (brain or computer) to the things in the world with which the system interacts – including, preeminently, the active, sentient body whose behavior the system controls. But Searle insists that these are not the relevant causal properties. He concedes the possibility in principle of duplicating the input-output competence of a human brain with a "formal program," which (suitably attached) would guide a body through the world exactly as that body's brain would, and thus would acquire all the relevant extra systemic causal properties of the brain. But such a brain substitute would utterly fail to produce intentionality in the process, Searle holds, becuase it would lack some other causal properties of the brain's internal operation.[1]

How, though, would we know that it lacked these properties, if all we knew was that it was (an implementation of) a formal program? Since Searle concedes that the operation of anything – and hence a human brain – can be described in terms of the execution of a formal program, the mere existence of such a level of description of a system would not preclude its having intentionality. It seems that it is only when we can see that the system in question is *only* the implementation of a formal program that we can conclude that it doesn't make a little intentionality on the side. But nothing could be only the implementation of a formal program; computers exude heat and noise in the course of their operations – why not intentionality too?

Besides, which is the major product and which the byproduct? Searle can hardly deny that brains do in fact produce lots of reliable and appropriate bodily control. They do this, he thinks, by producing intentionality, but he concedes that something – such as a computer with the right input-output rules – could produce the control without making or using any intentionality. But then control is the main product and intentionality just one (no doubt natural) means of obtaining it. Had our ancestors been nonintentional mutants with mere control systems, nature would just as readily have selected them instead. (I owe this point to Bob Moore.) Or, to look at the other side of the coin, brains with lots of intentionality but no control competence would be producers of an ecologically irrelevant product, which evolution would not protect. Luckily for us, though, our brains make intentionality; if they didn't, we'd behave just as we now do, but of course we wouldn't *mean* it!

Surely Searle does not hold the view I have just ridiculed, although it seems as if he does. He can't really view intentionality as a marvelous mental fluid, so what is he trying to get at? I think his concern with *internal* properties of control systems is a misconceived attempt to capture the interior *point of view* of a conscious agent. He does not see how any mere computer, chopping away at a formal program, could harbor such a point of view. But that is because he is looking *too deep*. It is just as mysterious if we peer into the synapse-filled jungles of the brain and wonder where consciousness is hiding. It is not at that level of description that a proper subject of consciousness will be found. That is the systems reply, which Searle does not yet see to be a step in the right direction away from his updated version of *élan vital*.

### Note

**1.** For an intuition pump involving exactly this case – a prosthetic brain – but designed to pump contrary intuitions, see "Where Am I?" in Dennett (1978).

*by* **John C. Eccles**

*Cá a lá Gra, Contra (Locarno) CH-6611, Switzerland*

## A dualist-interactionist perspective

Searle clearly states that the basis of his critical evaluation of AI is dependent on two propositions. The first is: "Intentionality in human beings (and animals) is a product of causal features of the brain." He supports this proposition by an unargued statement that it "is an empirical fact about the actual causal relations between mental processes and brains. It says simply that certain brain processes are *sufficient* for intentionality" (my italics).

This is a dogma of the psychoneural identity theory, which is one variety of the materialist theories of the mind. There is no mention of the alternative hypothesis of dualist interactionism that Popper and I published some time ago (1977) and that I have further developed more recently (Eccles 1978; 1979). According to that hypothesis intentionality is a property of the self-conscious mind (World 2 of Popper), the brain being used as an instrument in the realization of intentions. I refer to Fig. E 7-2 of Popper and Eccles (1977), where intentions appear in the box (inner senses) of World 2, with arrows indicating the flow of information by which intentions in the mind cause changes in the liaison brain and so eventually in voluntary movements.

I have no difficulty with proposition 2, but I would suggest that 3, 4, and 5 be rewritten with "mind" substituted for "brain." Again the statement: "*only* a machine could think, and only very special kinds of machines . . . with internal causal powers equivalent to those of brains" is the identity theory dogma. I say dogma becuase it is unargued and without empirical support. The identity theory is very weak empirically, being merely a theory of promise.

So long as Searle speaks about human performance without regarding intentionality as a property of the brain, I can appreciate that he has produced telling arguments against the strong AI theory. The story of the hamburger with the *Gedankenexperiment* of the Chinese symbols is related to Premack's attempts to teach the chimpanzee Sarah a primitive level of human language as expressed in symbols [See Premack:"Does the Chimpanzee Have a Theory of Mind?" *BBS* 1(4) 1978]. The criticism of Lenneberg (1975) was that, by conditioning, Sarah had learnt a symbol game, using symbols instrumentally, but had no idea that it was related to human language. He trained high school students with the procedures described by Premack, closely replicating Premack's study. The human subjects were quickly able to obtain considerably lower error scores than those reported for the chimpanzee. However, they were unable to translate correctly a single one of their completed sentences into English. In fact, they did not understand that there was any correspondence between the plastic symbols and language; instead they were under the impression that their task was to solve puzzles.

I think this simple experiment indicates a fatal flaw in all the AI work. No matter how complex the performance instantiated by the computer, it can be no more than a triumph for the computer designer in simulation. The Turing machine is a magician's dream – or nightmare!

It was surprising that after the detailed brain-mind statements of the abstract, I did not find the word "brain" in Searle's text through the whole of his opening three pages of argument, where he uses mind, mental states, human understanding, and cognitive states exactly as would be done in a text on dualist interactionism. Not until "the robot reply" does brain appear as "computer 'brain.' " However, from "the brain simulator reply" in the statements and criticisms of the various other replies, brain, neuron firings, synapses, and the like are profusely used in a rather naive way. For example "imagine the computer programmed with all the synapses of a human brain" is more than I can do by many orders of magnitude! So "the combination reply" reads like fantasy – and to no purpose!

I agree that it is a mistake to confuse simulation with duplication. But I do not object to the idea that the distinction between the program and its realization in the hardware seems to be parallel to the distinction between the mental operations and the level of brain operations. However, Searle believes that the equation "mind is to brain as program is to hardware" breaks down at several points. I would prefer to substitute programmer for program, because as a dualist interactionist I accept the analogy that as conscious beings we function as programmers of our brains. In particular I regret Searle's third argument: "Mental states and events are literally a product of the operation of the brain, but the program is not in that way a product of the computer," and so later we are told "whatever else intentionality is, it is a biological phenomenon, and it is as likely to be causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomenon." I have the feeling of being transported back to the nineteenth century, where, as derisorily recorded by Sherrington (1950): "the oracular Professor Tyndall, presiding over the British Association at Belfast, told his audience that as the bile is a secretion of the liver, so the mind is a secretion of the brain."

In summary, my criticisms arise from fundamental differences in

respect of beliefs in relation to the brain-mind problem. So long as Searle is referring to human intentions and performances without reference to the brain-mind problem, I can appreciate the criticisms that he marshals against the AI beliefs that an appropriately programmed computer is a mind literally understanding and having other cognitive states. Most of Searle's criticisms are acceptable for dualist interactionism. It is high time that strong AI was discredited.

*by* **J. A. Fodor**

*Department of Psychology, Massachusetts Institute of Technology, Cambridge, Mass. 02139*

## Searle on what only brains can do

1. Searle is certainly right that instantiating the same program that the brain does is not, in and of itself, a sufficent condition for having those propositional attitudes characteristic of the organism that has the brain. If some people in AI think that it is, they're wrong. As for the Turing test, it has all the usual difficulties with predictions of "no difference"; you can't distinguish the truth of the prediction from the insensitivity of the test instrument.[1]

2. However, Searle's treatment of the "robot reply" is quite unconvincing. Given that there are the right kinds of causal linkages between the symbols that the device manipulates and things in the world – including the afferent and efferent transducers of the device – it is quite unclear that intuition rejects ascribing propositional attitudes to it. All that Searle's example shows is that the kind of causal linkage he imagines – one that is, in effect, mediated by a man sitting in the head of a robot – is, unsurprisingly, not the right kind.

3. We don't know how to say what the right kinds of causal linkage are. This, also, is unsurprising since we don't know how to answer the closely related question as to what kinds of connection between a formula and the world determine the interpretation under which the formula is employed. We don't have an answer to this question for *any* symbolic system; a fortiori, not for mental representations. These questions are closely related because, given the mental representation view, it is natural to assume that what makes mental states intentional is primarily that they involve relations to semantically interpreted mental objects; again, relations of the right kind.

4. It seems to me that Searle has misunderstood the main point about the treatment of intentionality in representational theories of the mind; this is not surprising since proponents of the theory – especially in AI – have been notably unlucid in expounding it. For the record, then, the main point is this: intentional properties of propositional attitudes are viewed as inherited from semantic properties of mental representations (and not from the functional role of mental representations, unless "functional role" is construed broadly enough to include symbol-world relations). In effect, what is proposed is a reduction of the problem *what makes mental states intentional* to the problem *what bestows semantic properties on (fixes the interpretation of) a symbol.* This reduction looks promising because we're going to have to answer the latter question anyhow (for example, in constructing theories of natural languages); and we need the notion of mental representation anyhow (for example, to provide appropriate domains for mental processes).

It may be worth adding that there is nothing new about this strategy. Locke, for example, thought (a) that the intentional properties of mental states are inherited from the semantic (referential) properties of mental representations; (b) that mental processes are formal (associative); and (c) that the objects from which mental states inherit their intentionality are the same ones over which mental processes are defined: namely ideas. It's my view that no serious alternative to this treatment of propositional attitudes has ever been proposed.

5. To say that a computer (or a brain) performs formal operations on symbols is not the same thing as saying that it performs operations on formal (in the sense of "uninterpreted") symbols. This equivocation occurs repeatedly in Searle's paper, and causes considerable confusion. If there are mental representations they must, of course, be interpreted objects; it is because they are interpreted objects that mental states are intentional. But the brain might be a computer for all that.

6. This situation – needing a notion of causal connection, but not knowing which notion of causal connection is the right one – is entirely familiar in philosophy. It is, for example, extremely plausible that "a perceives b" can be true only where there is the right kind of causal connection between a and b. And we don't know what the right kind of causal connection is here either.

Demonstrating that some kinds of causal connection are the *wrong* kinds would not, of course, prejudice the claim. For example, suppose we interpolated a little man between a and b, whose function it is to report to a on the presence of b. We would then have (inter alia) a sort of causal link from a to b, but we wouldn't have the sort of causal link that is required for a to perceive b. It would, of course, be a fallacy to argue from the fact that this causal linkage fails to reconstruct perception to the conclusion that *no* causal linkage would succeed. Searle's argument against the "robot reply" is a fallacy of precisely that sort.

7. It is entirely reasonable (indeed it must be true) that the right kind of causal relation is the kind that holds between our brains and our transducer mechanisms (on the one hand) and between our brains and distal objects (on the other). It would not begin to follow that *only* our brains can bear such relations to transducers and distal objects; and it would also not follow that being the same sort of thing our brain is (in any biochemical sense of "same sort") is a necessary condition for being in that relation; and it would also not follow that formal manipulations of symbols are not among the links in such causal chains. And, even if our brains *are* the only sorts of things that can be in that relation, the fact that they are might quite possibly be of no particular interest; that would depend on *why* it's true.[2]

Searle gives no clue as to why he thinks the biochemistry is important for intentionality and, prima facie, the idea that what counts is how the organism is connected to the world seems far more plausible. After all, it's easy enough to imagine, in a rough and ready sort of way, how the fact that my thought is causally connected to a tree might bear on its being a thought about a tree. But it's hard to imagine how the fact that (to put it crudely) my thought is made out of hydrocarbons could matter, except on the unlikely hypothesis that only hydrocarbons can be causally connected to trees in the way that brains are.

8. The empirical evidence for believing that "manipulation of symbols" is involved in mental processes derives largely from the considerable success of work in linguistics, psychology, and AI that has been grounded in that assumption. Little of the relevant data concerns the simulation of behavior or the passing of Turing tests, though Searle writes as though all of it does. Searle gives no indication *at all* of how the facts that this work accounts for are to be explained if not on the mental-processes-are-formal-processes view. To claim that there is no argument that symbol manipulation is necessary for mental processing while systematically ignoring all the evidence that has been alleged in favor of the claim strikes me as an extremely curious strategy on Searle's part.

9. Some necessary conditions are more interesting than others. While connections to the world and symbol manipulations are both presumably necessary for intentional processes, there is no reason (so far) to believe that the former provide a theoretical domain for a science; wheras, there is considerable a posteriori reason to suppose that the latter do. If this is right, it provides some justification for AI practice, if not for AI rhetoric.

10. *Talking* involves performing certain formal operations on symbols: stringing words together. Yet, not everything that can string words together can talk. It does not follow from these banal observations that what we utter are uninterpreted sounds, or that we don't understand what we say, or that whoever talks talks nonsense, or that only hydrocarbons can assert – similaly, mutatis mutandis, if you substitue "thinking" for "talking."

### Notes

**1.** I assume, for simplicity, that there is only one program that the brain instantiates (which, of course, there isn't). Notice, by the way, that even passing the Turing test requires doing more than *just* manipulating symbols. A device that can't run a typewriter can't play the game.

**2.** For example, it might be that, in point of physical fact, only things that have the same simultaneous values of weight, density, and shade of gray that brains have can do the things that brains can. This would be surprising, but it's hard to see why a psychologist should care much. Not even if it turned out – still in point of physical fact – that brains are the only things that *can* have that weight, density, and color. If *that's* dualism, I imagine we can live with it.

*by* **John Haugeland**

*Center for Advanced Study in the Behavioral Sciences, Stanford, Calif. 94305*

## Programs, causal powers, and intentionality

Searle is in a bind. He denies that any Turing test for intelligence is adequate – that is, that behaving intelligently is a sufficent condition for being intelligent. But he dare not deny that creatures physiologically very different from people might be intelligent nonetheless – smart green saucer pilots, say. So he needs an intermediate criterion: not so specific to us as to rule out the aliens, yet not so dissociated from specifics as to admit any old object with the right behavior. His suggestion is that only objects (made of stuff) with "the right causal powers" can have intentionality, and hence, only such objects can genuinely understand anything or be intelligent. This suggestion, however, is incompatible with the main argument of his paper.

Ostensibly, that argument is against the claim that working according to a certain program can ever be sufficient for understanding anything – no matter how cleverly the program is contrived so as to make the relevant object (computer, robot, or whatever) behave *as if* it understood. The crucial move is replacing the central processor (c.p.u.) with a superfast person – whom we might as well call "Searle's demon." And Searle argues that an English-speaking demon could perfectly well follow a program for simulating a Chinese speaker, without itself understanding a word of Chinese.

The trouble is that the same strategy will work as well against any specification of "the right causal powers." Instead of manipulating formal tokens according to the specifications of some computer program, the demon will manipulate physical states or variables according to the specification of the "right" causal interactions. Just to be concrete, imagine that the right ones are those powers that our neuron tips have to titillate one another with neurotransmitters. The green aliens can be intelligent, even though they're based on silicon chemistry, because their (silicon) neurons have the same power of intertitillation. Now imagine covering each of the neurons of a Chinese criminal with a thin coating, which has no effect, except that it is impervious to neurotransmitters. And imagine further that Searle's demon can see the problem, and comes to the rescue; he peers through the coating at each neural tip, determines which transmitter (if any) would have been emitted, and then massages the adjacent tips in a way that has the same effect as if they had received that transmitter. Basically, instead of replacing the c.p.u., the demon is replacing the neurotransmitters.

By hypothesis, the victim's behavior is unchanged; in particular, she still acts as if she understood Chinese. Now, however, none of her neurons has the right causal powers – the demon has them, and he still understands only English. Therefore, having the right causal powers (even while embedded in a system such that the exercise of these powers leads to "intelligent" behavior) cannot be sufficient for understanding. Needless to say, a corresponding variation will work, whatever the relevant causal powers are.

None of this should come as a surprise. A computer program just is a specification of the exercise of certain causal powers: the powers to manipulate various formal tokens (physical objects or states of some sort) in certain specified ways, depending on the presence of certain other such tokens. Of course, it is a particular way of specifying causal exercises of a particular sort – that's what gives the "computational paradigm" its distinctive character. But Searle makes no use of this particularity; his argument depends *only* on the fact that causal powers can be specified independently of whatever it is that has the power. This is precisely what makes it possible to interpose the demon, in both the token-interaction (program) and neuron-interaction cases.

There is no escape in urging that this is a "dualistic" view of causal powers, not intrinisically connected with "the actual properties" of physical objects. To speak of causal powers in any way that allows for generalization (to green aliens, for example) is ipso facto to abstract from the particulars of any given "realization." The point is independent of the example – it works just as well for photosynthesis. Thus, flesh-colored plantlike organisms on the alien planet might photosynthesize (I take it, in a full and literal sense) so long as they contain some chemical (not necessarily chlorophyll) that absorbs light and uses the energy to make sugar and free oxygen out of carbon dioxide (or silicon dioxide?) and water. This is what it means to specify photosynthesis as a *causal power*, rather than just a property that is, by definition, idiosyncratic to chlorophyll. But now, of course, the demon can enter, replacing both chlorophyll and its alien substitute: he devours photons, and thus energized, makes sugar from $CO_2$ and $H_2O$. It seems to me that the demon is photosynthesizing.

Let's set aside the demon argument, however. Searle also suggests that "there is no reason to suppose" that understanding (or intentionality) "has anything to do with" computer programs. This too, I think, rests on his failure to recognize that specifying a program is (in a distinctive way) specifying a range of causal powers and interactions.

The central issue is what differentiates *original* intentionality from *derivative* intentionality. The former is intentionality that a thing (system, state, process) has "in its own right"; the latter is intentionality that is "borrowed from" or "conferred by" something else. Thus (on standard assumptions, which I will not question here), the intentionality of conscious thought and perception is original, whereas the intentionality (meaning) of linguistic tokens is merely conferred upon them by language users – that is, words don't have any meaning in and of themselves, but only in virtue of our giving them some. These are paradigm cases; many other cases will fall clearly on one side or the other, or be questionable, or perhaps even marginal. No one denies that if AI systems don't have original intentionality, then they at least have derivative intentionality, in a nontrivial sense – because they have nontrivial *interpretations*. What Searle objects to is the thesis, held by many, that good-enough AI systems have (or will eventually have) original intentionality.

Thought tokens, such as articulate beliefs and desires, and linguistic tokens, such as the expressions of articulate beliefs and desires, seem to have a lot in common – as pointed out, for example, by Searle (1979c). In particular, except for the original/derivative distinction, they have (or at least appear to have) closely parallel semantic structures and variations. There must be some other principled distinction between them, then, in virtue of which the former can be originally intentional, but the latter only derivatively so. A conspicuous candidate for this distinction is that thoughts are semantically active, whereas sentence tokens, written out, say, on a page, are semantically inert. Thoughts are constantly interacting with one another and the world, in ways that are semantically appropriate to their intentional content. The causal interactions of written sentence tokens, on the other hand, do not consistently reflect their content (except when they interact with people).

Thoughts are embodied in a "system" that provides "normal channels" for them to interact with the world, and such that these normal interactions tend to maximize the "fit" between them and the world; that is, via perception, beliefs tend toward the truth; and, via action, the world tends toward what is desired. And there are channels of interaction among thoughts (various kinds of inference) via which the set of them tends to become more coherent, and to contain more consequences of its members. Naturally, other effects introduce aberrations and "noise" into the system; but the normal channels tend to predominate in the long run. There are no comparable channels of interaction for written tokens. In fact, (according to this same standard view), the only semantically sensitive interactions that written tokens ever have are with thoughts; insofar as they tend to express truths, it is because they express beliefs, and insofar as they tend to bring about their own satisfaction conditions, it is because they tend to bring about desires. Thus, the *only* semantically significant interactions that written tokens have with the world are via thoughts; and this, the suggestion

goes, is *why* their intentionality is derivative.

The interactions that thoughts have among themselves (within a single "system") are particularly important, for it is in virtue of these that thought can be subtle and indirect, relative to its interactions with the world – that is, not easily fooled or thwarted. Thus, we tend to consider more than the immediately present evidence in making judgments, and more than the immediately present options in making plans. We weigh desiderata, seek further information, try things to see if they'll work, formulate general maxims and laws, estimate results and costs, go to the library, cooperate, manipulate, scheme, test, and reflect on what we're doing. All of these either are or involve a lot of thought-thought interaction, and tend, in the long run to broaden and improve the "fit" between thought and world. And they are typical as manifestations both of intelligence and of independence.

I take it for granted that all of the interactions mentioned are, in some sense, *causal* – hence, that it is among the system's "causal powers" that it can have (instantiate, realize, produce) thoughts that interact with the world and each other in these ways. It is hard to tell whether these are the sorts of causal powers that Searle has in mind, both because he doesn't say, and because they don't seem terribly similar to photosynthesis and lactation. But, in any case, they strike me as strong candidates for the kinds of powers that would distinguish systems with intentionality – that is, *original* intentionality – from those without. The reason is that these are the only powers that consistently reflect the distinctively intentional character of the interactors: namely, their "content" or "meaning" (except, so to speak, passively, as in the case of written tokens being read). That is, the power to have states that are semantically active is the "right" causal power for intentionality.

It is this plausible claim that underlies the thesis that (sufficiently developed) AI systems could actually *be* intelligent, and have *original* intentionality. For a case can surely be made that their "representations" are semantically active (or, at least, that they would be if the system were built into a robot). Remember, we are conceding them at least derivative intentionality, so the states in question do have a content, relative to which we can gauge the "semantic appropriateness" of their causal interactions. And the central discovery of all computer technology is that devices can be contrived such that, relative to a certain interpretation, certain of their states will always interact (causally) in semantically appropriate ways, so long as the devices perform as designed electromechanically – that is, these states can have "normal channels" of interaction (with each other and with the world) more or less comparable to those that underlie the semantic activity of thoughts. This point can hardly be denied, so long as it is made in terms of the derivative intentionality of computing systems; but what it seems to add to the archetypical (and "inert") derivative intentionality of, say, written text is, precisely, semantic *activity*. So, if (sufficiently rich) semantic activity is what distinguishes original from derivative intentionality (in other words, it's the "right" causal power), then it seems that (sufficiently rich) computing systems can have *original* intentionality.

Now, like Searle, I am inclined to dispute this conclusion; but for entirely different reasons. I don't believe there is any *conceptual* confusion in supposing that the right causal powers for original intentionality are the ones that would be captured by specifying a program (that is, a virtual machine). Hence, I don't think the above plausibility argument can be dismissed out of hand ("no reason to suppose," and so on); nor can I imagine being convinced that, no matter how good AI got, it would still be "weak" – that is, would not have created a "real" intelligence – because it still proceeded by specifying programs. It seems to me that the interesting question is much more nitty-gritty empirical than that: given that programs *might* be the right way to express the relevant causal structure, are they *in fact* so? It is to this question that I expect the answer is no. In other words, I don't much care about Searle's demon working through a program for perfect simulation of a native Chinese speaker – not because there's no such demon, but because there's no such program. Or rather, *whether* there is such a program, and *if not, why not*, are, in my view, the important questions.

*by* **Douglas R. Hofstadter**

*Computer Science Department, Indiana University, Bloomington, Ind. 47405*

## Reductionism and religion

This religious diatribe against AI, masquerading as a serious scientific argument, is one of the wrongest, most infuriating articles I have ever read in my life. It is matched in its power to annoy only by the famous article "Minds, Machines, and Gödel" by J. R. Lucas (1961).

Searle's trouble is one that I can easily identify with. Like me, he has deep difficulty in seeing how mind, soul, "I," can come out of brain, cells, atoms. To show his puzzlement, he gives some beautiful paraphrases of this mystery. One of my favorites is the water-pipe simulation of a brain. It gets straight to the core of the mind-body problem. The strange thing is that Searle simply dismisses any possibility of such a system's being conscious with a hand wave of "absurd." (I actually think he radically misrepresents the complexity of such a water-pipe system both to readers and in his own mind, but that is a somewhat separable issue.)

The fact is, we have to deal with a reality of nature – and realities of nature sometimes *are* absurd. Who would have believed that light consists of spinning massless wave particles obeying an uncertainty principle while traveling through a curved four-dimensional universe? The fact that intelligence, understanding, mind, consciousness, soul all do spring from an unlikely source – an enormously tangled web of cell bodies, axons, synapses, and dendrites – is absurd, and yet undeniable. How this can create an "I" is hard to understand, but once we accept that fundamental, strange, disorienting fact, then it should seem no more weird to accept a water-pipe "I."

Searle's way of dealing with this reality of nature is to claim he accepts it – but then he will not accept its consequences. The main consequence is that "intentionality" – his name for soul – is an outcome of formal processes. I admit that I have slipped one extra premise in here: that physical processes are formal, that is, rule governed. To put it another way, the extra premise is that there is no intentionality at the level of particles. (Perhaps I have misunderstood Searle. He may be a mystic and claim that there is intentionality at that level. But then how does one explain why it seems to manifest itself in consciousness only when the particles are arranged in certain special configurations – brains – but not, say, in water-pipe arrangements of any sort and size?) The conjunction of these two beliefs seems to me to compel one to admit the possibility of all the hopes of artificial intelligence, despite the fact that it will always baffle us to think of ourselves as, at bottom, formal systems.

To people who have never programmed, the distinction between levels of a computer system – programs that run "on" other programs or on hardware – is an elusive one. I believe Searle doesn't really understand this subtle idea, and thus blurs many distinctions while creating other artificial ones to take advantage of human emotional responses that are evoked in the process of imagining unfamiliar ideas.

He begins with what sounds like a relatively innocent situation: a man in a room with a set of English instructions ("bits of paper") for manipulating some Chinese symbols. At first, you think the man is answering questions (although unbeknown to him) about restaurants, using Schankian scripts. Then Searle casually slips in the idea that this *program can pass the Turing test! This is an incredible jump in* complexity – perhaps a millionfold increase if not more. Searle seems not to be aware of how radically it changes the whole picture to have that "little" assumption creep in. But even the initial situation, which sounds plausible enough, is in fact highly unrealistic.

Imagine a human being, hand simulating a complex AI program, such as a script-based "understanding" program. To digest a full story, to go through the scripts and to produce the response, would probably take a hard eight-hour day for a human being. Actually, of course, this hand-simulated program is supposed to be passing the Turing test, not just answering a few stereotyped questions about restaurants. So let's jump up to a week per question, since the program would be so complex. (We are being unbelievably generous to Searle.)

Now Searle asks you to identify with this poor slave of a human (he doesn't actually ask you to identify with him – he merely knows you will project yourself onto this person, and vicariously experience the indescribably boring nightmare of that hand simulation). He knows your reaction will be: "This is not understanding the story – this is some sort of formal process!" But remember: any time some phenomenon is looked at on a scale a million times different from its familiar scale, it doesn't seem the same! When I imagine myself feeling my brain running a hundred times too slowly (of course that is paradoxical but it is what Searle wants me to do), then of course it is agonizing, and presumably I would not even recognize the feelings at all. Throw in yet another factor of a thousand and one cannot even imagine what it would feel like.

Now this is what Searle is doing. He is inviting you to identify with a nonhuman which he lightly passes off as a human, and by doing so he asks you to participate in a great fallacy. Over and over again he uses this ploy, this emotional trickery, to get you to agree with him that surely, an intricate system of water pipes can't think! He forgets to tell you that a water-pipe simulation of the brain would take, say, a few trillion water pipes with a few trillion workers standing at faucets turning them when needed, and he forgets to tell you that to answer a question it would take a year or two. He forgets to tell you, because if you remembered that, and then on your own, imagined taking a movie and speeding it up a million times, and imagined changing your level of description of the thing from the faucet level to the pipe-cluster level, and on through a series of ever higher levels until you reached some sort of eventual symbolic level, why then you might say, "Hey, when I imagine what this entire system would be like when perceived at this time scale and level of description, I can see how it might be conscious after all!"

Searle is representative of a class of people who have an instinctive horror of any "explaining away" of the soul. I don't know why certain people have this horror while others, like me, find in reductionism the ultimate religion. Perhaps my lifelong training in physics and science in general has given me a deep awe at seeing how the most substantial and familiar of objects or experiences fades away, as one approaches the infinitesimal scale, into an eerily insubstantial ether, a myriad of ephemeral swirling vortices of nearly incomprehensible mathematical activity. This in me evokes a kind of cosmic awe. To me, reductionism does not "explain away"; rather, it adds mystery. I know that this journal is not the place for philosophical and religious commentary, yet it seems to me that what Searle and I have is, at the deepest level, a religious disagreement, and I doubt that anything I say could ever change his mind. He insists on things he calls "causal intentional properties" which seem to vanish as soon as you analyze them, find rules for them, or simulate them. But what those things are, other than epiphenomena, or "innocently emergent" qualities, I don't know.

*by* **B. Libet**

*Department of Physiology, University of California, San Francisco, Calif. 94143*

## Mental phenomena and behavior

Searle states that the main argument of his paper is directed at establishing his second proposition, that "instantiating a computer program is never by itself a sufficient condition of intentionality" (that is, of a mental state that includes beliefs, desires, and intentions). He accomplishes this with a *Gedankenexperiment* to show that even "a human agent could instantiate the program and still not have the relevant intentionality"; that is, Searle shows, in a masterful and convincing manner, that the behavior of the appropriately programmed computer could transpire in the absence of a cognitive mental state. I believe it is also possible to establish the proposition by means of an argument based on simple formal logic.

We start with the knowledge that we are dealing with two different systems: system A is the computer, with its appropriate program; system B is the human being, particularly his brain. Even if system A could be arranged to behave and even to look like system B, in a manner that might make them indistinguishable to an external observer, system A must be at least internally different from B. If A and B were identical, they would both be human beings and there would be no thesis to discuss.

Let us accept the proposal that, on an input-output basis, system A and system B could be made to behave alike, properties that we may group together under category X. The possession of the relevant mental states (including understanding, beliefs, desires, intentions, and the like) may be called property Y. We know that system B has property Y. Remembering that systems A and B are known to be different, it is an error in logic to argue that because systems A and B both have property X, they must also both have property Y.

The foregoing leads to a more general proposition – that no behavior of a computer, regardless of how successful it may be in simulating human behavior, is ever by itself sufficient evidence of any mental state. Indeed, Searle also appears to argue for this more general case when, later in the discussion, he notes: (a) To get computers to feel pain or fall in love would be neither harder nor easier than to get them to have cognition. (b) "For simulation, all you need is the right input and output and a program in the middle that transforms the former into the latter." And, (c) "to confuse simulation with duplication is the same mistake, whether it is pain, love, cognition." On the other hand, Searle seems not to maintain this general proposition with consistency. In his discussion of "IV. The combination reply" (to his analytical example or thought experiment), Searle states: "If we could build a robot whose behavior was indistinguishable over a large range from human behavior, we would . . . find it rational and indeed irresistible to . . . attribute intentionality to it, pending some reason not to." On the basis of my argument, one would not have to know that the robot had a formal program (or whatever) that accounts for its behavior, in order not to have to attribute intentionality to it. All we need to know is that the robot's internal control apparatus is not made in the same way and out of the same stuff as is the human brain, to reject the thesis that the robot must possess the mental states of intentionality, and so on.

Now, it is true that neither my nor Searle's argument excludes the possibility that an appropriately programmed computer could also have mental states (property Y); the argument merely states it is not warranted to propose that the robot must have mental states (Y). However, Searle goes on to contribute a valuable analysis of why so many people have believed that computer programs do impart a kind of mental process or state to the computer. Searle notes that, among other factors, a residual behaviorism or operationalism underlies the willingness to accept input-output patterns as sufficient for postulating human mental states in appropriately programmed computers. I would add that there are still many psychologists and perhaps philosophers who are similarly burdened with residual behaviorism or operationalism even when dealing with criteria for existence of a conscious subjective experience in human subjects (see Libet 1973; 1979).

*by* **William G. Lycan**

*Department of Philosophy, Ohio State University, Columbus, Ohio 43210*

## The functionalist reply (Ohio State)

Most versions of philosophical behaviorism have had the consequence that if an organism or device *D* passes the Turing test, in the sense of systematically manifesting all the same outward behavioral dispositions that a normal human does, the *D* has all the same sorts of contentful or intentional states that humans do. In light of fairly obvious counterexamples to this thesis, materialist philosophers of mind have by and large rejected behaviorism in favor of a more species-chauvinistic view: *D's* manifesting all the same sorts of behavioral dispositions we do does not alone suffice for *D's* having intentional states; it is necessary in addition that *D* produce behavior from stimuli *in roughly the way that* we do – that *D's* inner functional organization be not unlike ours and that *D* process the stimulus input by analogous inner procedures. On this "functionalist" theory, to be in a mental state of such and such a kind is to incorporate a functional component or system of components of type so and so which is in a certain distinctive state of its own. "Functional components" are individuated according to the roles they play within their owners' overall functional organization.[1]

Searle offers a number of cases of entities that manifest the

behavioral dispositions we associate with intentional states but that rather plainly do not have any such states.[2] I accept his intuitive judgments about most of these cases. Searle plus rule book plus pencil and paper presumably does not understand Chinese, nor does Searle with memorized rule book or Searle with TV camera or the robot with Searle inside. Neither my stomach nor Searle's liver nor a thermostat nor a light switch has beliefs and desires. But none of these cases is a counterexample to the functionalist hypothesis. The systems in the former group are pretty obviously not functionally isomorphic at the relevant level to human beings who do understand Chinese; a native Chinese carrying on a conversation is implementing procedures of his own, not those procedures that would occur in a mockup containing the cynical, English-speaking, American-acculturated homuncular Searle. Therefore they are not counterexamples to a functionalist theory of language understanding, and accordingly they leave it open that a computer that *was* functionally isomorphic to a real Chinese speaker would indeed understand Chinese also. Stomachs, thermostats, and the like, because of their brutish simplicity, are even more clearly dissimilar to humans. (The same presumably is true of Schank's existing language-understanding programs.)

I have hopes for a sophisticated version of the "brain simulator" (or the "combination" machine) that Searle illustrates with his plumbing example. Imagine a hydraulic system of this type that does replicate, perhaps not the precise neuroanatomy of a Chinese speaker, but all that is relevant of the Chinese speaker's higher functional organization; individual water pipes are grouped into organ systems precisely analogous to those found in the speaker's brain, and the device processes linguistic input in just the way that the speaker does. (It does not merely *simulate* or *describe* this processing.) Moreover, the system is automatic and does all this without the intervention of Searle or any other *deus in machina*. Under these conditions and given a suitable social context, I think it would be plausible to accept the functionalist consequence that the hydraulic system does understand Chinese.

Searle's paper suggest two objections to this claim. First, "where is the understanding in this system?" All Searle sees is pipes and valves and flowing water. *Reply:* Looking around the fine detail of the system's hardware, you are *too small* to see that the system is understanding Chinese sentences. If you were a tiny, cell-sized observer inside a real Chinese speaker's brain, all you would see would be neurons stupidly, mechanically transmitting electrical charge, and in the same tone you would ask, "Where is the understanding in this system?" But you would be wrong in concluding that the system you were observing did not understand Chinese; in like manner you may well be wrong about the hydraulic device.[3]

Second, even if a computer were to replicate all of the Chinese speaker's relevant functional organization, all the computer is really doing is performing computational operations on formally specified elements. A purely formally or syntactically characterized element has no meaning or content in itself, obviously, and no amount of mindless syntactic manipulation of it will endow it with any. *Reply:* The premise is correct, and I agree it shows that no computer has or could have intentional states *merely in virtue of performing syntactic operations on formally characterized elements.* But that does not suffice to prove that no computer can have intentional states at all. Our brain states do not have the contents they do just in virtue of having their purely formal properties either;[4] a brain state described "syntactically" has no meaning or content on its own. In virtue of what, then, *do* brain states (or mental states however construed) have the meanings they do? Recent theory advises that the content of a mental representation is not determined within its owner's head (Putnam 1975a; Fodor 1980); rather, it is determined in part by the objects in the environment that actually figure in the representation's etiology and in part by social and contextual factors of several other sorts (Stich, in preparation). Now, present-day computers live in highly artificial and stifling environments. They receive carefully and tendentiously preselected input; their software is adventitiously manipulated by uncaring programmers; and they are isolated in laboratories and offices, deprived of any normal interaction within a natural or appropriate social setting.[5] For this reason and several others, Searle is surely right in saying that present-day computers do not really have the intentional states that we fancifully incline toward attributing to them. But nothing Searle has said impugns the thesis that if a sophisticated future computer not only replicated human functional organization but harbored its inner representations as a result of the right sort of causal history and had also been nurtured within a favorable social setting, we might correctly ascribe intentional states to it. This point may or may not afford lasting comfort to the AI community.

### *Notes*

**1.** This characterization is necessarily crude and vague. For a very useful survey of different versions of functionalism and their respective foibles, see Block (1978); I have developed and defended what I think is the most promising version of functionalism in Lycan (forthcoming).

**2.** For further discussion of cases of this kind, see Block (forthcoming).

**3.** A much expanded version of this reply appears in section 4 of Lycan (forthcoming).

**4.** I do not understand Searle's positive suggestion as to the source of intentionality in our own brains. *What* "neurobiological causal properties"?

**5.** As Fodor (forthcoming) remarks, SHRDLU as we interpret him is the victim of a Cartesian evil demon; the "blocks" he manipulates do not exist in reality.

*by* **John McCarthy**

*Artificial Intelligence Laboratory, Stanford University, Stanford, Calif. 94305*

## Beliefs, machines, and theories

John Searle's refutation of the Berkeley answer that the system understands Chinese proposes that a person (call him Mr. Hyde) carry out in his head a process (call it Dr. Jekyll) for carrying out a written conversation in Chinese. Everyone will agree with Searle that Mr. Hyde does not understand Chinese, but I would contend, and I suppose his Berkeley interlocutors would also, that provided certain other conditions for understanding are met, Dr. Jekyll understands Chinese. In Robert Louis Stevenson's story, it seems assumed that Dr. Jekyll and Mr. Hyde time-share the body, while in Searle's case, one interprets a program specifying the other.

Searle's dismissal of the idea that thermostats may be ascribed belief is based on a misunderstanding. It is not a pantheistic notion that all machinery including telephones, light switches, and calculators believe. Belief may usefully be ascribed only to systems about which someone's knowledge can best be expressed by ascribing beliefs that satisfy axioms such as those in McCarthy (1979). Thermostats are sometimes such systems. Telling a child, "If you hold the candle under the thermostat, you will fool it into thinking the room is too hot, and it will turn off the furnace" makes proper use of the child's repertoire of intentional concepts.

Formalizing belief requires treating simple cases as well as more interesting ones. Ascribing beliefs to thermostats is analogous to including 0 and 1 in the number system even though we would not need a number system to treat the null set or sets with just one element; indeed we wouldn't even need the concept of set.

However, a program that understands should not be regarded as a theory of understanding any more than a man who understands is a theory. A program can only be an illustration of a theory, and a useful theory will contain much more than an assertion that "the following program understands about restaurants." I can't decide whether this last complaint applies to Searle or just to some of the AI researchers he criticizes.

*by* **John C Marshall**

*Neuropsychology Unit, University Department of Clinical Neurology, The Radcliffe Infirmary, Oxford, England*

## Artificial intelligence—the real thing?

Searle would have us believe that the present-day inhabitants of respectable universities have succumbed to the Faustian dream. (Mephistopheles: "What is it then?" Wagner: "A man is in the making.") He assures us, with a straight face, that some contemporary scholars think that "an appropriately programmed computer really is a mind," that such artificial creatures "literally have cognitive states." The real thing indeed! But surely no one could believe this? I mean, if

someone did, then wouldn't he want to give his worn-out IBM a decent burial and say Kaddish for it? And even if some people at Yale, Berkeley, Stanford, and so forth do instantiate these weird belief states, what conceivable *scientific* interest could that hold? Imagine that they were right, and that their computers really do perceive, understand, and think. All that our Golem makers have done on Searle's story is to create yet another mind. If the sole aim is to "reproduce" (Searle's term, not mine) mental phenomena, there is surely no need to buy a computer.

Frankly, I just don't care what some members of the AI community think about the ontological status of their creations. What I do care about is whether anyone can produce principled, revealing accounts of, say, the perception of tonal music (Longuet-Higgins 1979), the properties of stereo vision (Marr & Poggio 1979), and the parsing of natural language sentences (Thorne 1968). Everyone that I know who tinkers around with computers does so because he has an attractive theory of some psychological capacity and wishes to explore certain consequences of the theory algorithmically. Searle refers to such activity as "weak AI," but I would have thought that theory construction and testing was one of the stronger enterprises that a scientist could indulge in. Clearly, there must be some radical misunderstanding here.

The problem appears to lie in Searle's (or his AI informants') strange use of the term 'theory.' Thus Searle writes in his shorter abstract: "According to strong AI, appropriately programmed computers literally have cognitive states, and therefore the programs are psychological theories." Ignoring momentarily the "and therefore," which introduces a simple non sequitur, how could a program be a theory? As Moor (1978) points out, a theory is (at least) a collection of related propositions which may be true or false, whereas a program is (or was) a pile of punched cards. For all I know, maybe suitably switched-on computers do "literally" have cognitive states, but even if they did, how could that possibly licence the inference that the program per se was a psychological theory? What would one make of an analogous claim applied to physics rather than psychology? "Appropriately programmed computers literally have physical states, and therefore the programs are theories of matter" doesn't sound like a valid inference to me. Moor's exposition of the distinction between program and theory is particularly clear and thus worth quoting at some length:

> A program must be interpreted in order to generate a theory. In the process of interpreting, it is likely that some of the program will be discarded as irrelevant since it will be devoted to the technicalities of making the program acceptable to the computer. Moreover, the remaining parts of the program must be organized in some coherent fashion with perhaps large blocks of the computer program taken to represent specific processes. Abstracting a theory from the program is not a simple matter, for different groupings of the program can generate different theories. Therefore, to the extent that a program, understood as a model, embodies one theory, it may well embody many theories. (Moor 1978, p. 215)

Searle reports that some of his informants believe that running programs *are* other minds, albeit artificial ones; if that were so, would these scholars not attempt to construct theories of artificial minds, just as we do for natural ones? Considerable muddle then arises when Searle's informants ignore their own claim and use the terms 'reproduce' and 'explain' synonymously: "The project is to reproduce and explain the mental by designing programs." One can see how hopelessly confused this is by transposing the argument back from computers to people. Thus I have noticed that many of my daughter's mental states bear a marked resemblance to my own; this has arisen, no doubt, because part of my genetic plan was used ot build her hardware and because I have shared in the responsibility of programming her. All well and good, but it would be straining credulity to regard my daughter as "explaining" me, as being a "theory" of me.

What one would like is an elucidation of the senses in which programs, computers and other machines do and don't figure in the explanation of behavior (Cummins 1977; Marshall 1977). It is a pity that Searle disregards such questions in order to discuss the everyday use of mental vocabulary, an enterprise best left to lexicographers. Searle writes: "The study of the mind starts with such facts as that humans have beliefs, while thermostats, telephones, and adding machines don't." Well, perhaps it does start there, but that is no reason to suppose it must finish there. How would such an "argument" fare in natural philosophy? "The study of physics starts with such facts as that tables are solid objects without holes in them, whereas Gruyere cheese. . . ." Would Searle now continue that "If you get a theory that denies this point, you have produced a counterexample to the theory and the theory is wrong"? Of course a thermostat's "belief" that the temperature should be a little higher is not the same kind of thing as my "belief" that it should. It would be totally uninteresting if they were the same. Surely the theorist who compares the two must be groping towards a deeper parallel; he has seen an analogy that may illuminate certain aspects of the control and regulation of complex systems. The notion of positive and negative feedback is what makes thermostats so appealing to Alfred Wallace and Charles Darwin, to Claude Bernard and Walter Cannon, to Norbert Wiener and Kenneth Craik. Contemplation of governors and thermostats has enabled them to see beyond appearances to a level at which there are profound similarities between animals and artifacts (Marshall 1977). It is Searle, not the theoretician, who doesn't really take the enterprise seriously. According to Searle, "what we wanted to know is what distinguishes the mind from thermostats and livers," Yes, but that is not all; we also want to know at what levels of description there are striking resemblances between disparate phenomena.

In the opening paragraphs of *Leviathan*, Thomas Hobbes (1651, p. 8) gives clear expression to the mechanist's philosophy:

> Nature, the art whereby God hath made and governs the world, is by the art of man, as in many other things, in this also imitated, that it can make an artificial animal. . . . For what is the *heart* but a *spring*, and the *nerves* so many *strings*; and *joints* so many *wheels* giving motion to the whole body, such as was intended by the artificer?

What is the notion of "imitation" that Hobbes is using here? Obviously not the idea of *exact* imitation or copying. No one would *confuse* a cranial nerve with a piece of string, a heart with the mainspring of a watch, or an ankle with a wheel. There is no question of *trompe l'oeil*. The works of the scientist are not in that sense *reproductions* of nature; rather they are attempts to see behind the phenomenological world to a hidden reality. It was Galileo, of course, who articulated this paradigm most forcefully: sculpture, remarks Galileo,

> is "closer to nature" than painting in that the material substratum manipulated by the sculptor shares with the matter manipulated by nature herself the quality of three-dimensionality. But does this fact rebound to the credit of sculpture? On the contrary, says Galileo, it greatly "diminishes its merit": What will be so wonderful in imitating sculptress Nature by sculpture itself?" And he concludes: "The most artistic imitation is that which represents the three-dimensional by its opposite, which is the plane." (Panofsky 1954, p. 97)

Galileo summarizes his position in the following words: "The further removed the means of imitation are from the thing to be imitated, the more worthy of admiration the imitation will be" (Panofsky 1954). In a footnote to the passage, Panofsky remarks on "the basic affinity between the spirit of this sentence and Galileo's unbounded admiration for Aristarchus and Copernicus 'because they trusted reason rather than sensory experience" (Panofsky 1954).

Now Searle is quite right in pointing out that in AI one seeks to model cognitive states and their consequences (the real thing) by a formal syntax, the interpretation of which exists only the eye of the beholder. Precisely therein lies the beauty and significance of the enterprise – to try to provide a counterpart for each substantive distinction with a syntactic one. This is essentially to regard the study of the relationships between physical transactions and symbolic operations as an essay in cryptanalysis (Freud 1895; Cummins 1977). The interesting question then arises as to whether there is a unique mapping between the formal elements of the system and their "meanings" (Householder 1962).

Searle, however, seems to be suggesting that we abandon entirely both the Galilean and the "linguistic" mode in order merely to copy cognitions. He would apparently have us seek mind only in "neurons with axons and dendrites," although he admits, as an empirical possibility, that such objects might "produce consciousness, intentionality and all the rest of it using some other sorts of chemical principles

than human beings use." But this admission gives the whole game away. How would Searle know that he had built a silicon-based mind (rather than our own carbon-based mind) except by having an appropriate *abstract* (that is, nonmaterial) characterization of what the two life forms hold in common? Searle finesses this problem by simply "attributing" cognitive states to himself, other people, and a variety of domestic animals: "In 'cognitive sciences' one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects." But this really won't do: we are, after all, a long way from having any very convincing evidence that cats and dogs have "cognitive states" in anything like Searle's use of the term [See "Cognition and Consciousness in Nonhuman Species" *BBS* 1(4) 1978].

Thomas Huxley (1874, p. 156) poses the question in his paraphrase of Nicholas Malebranche's orthodox Cartesian line: "What proof is there that brutes are other than a superior race of marionettes, which eat without pleasure, cry without pain, desire nothing, know nothing, and only simulate intelligence as a bee simulates a mathematician?" Descartes' friend and correspondent, Marin Mersenne, had little doubt about the answer to this kind of question. In his discussion of the perceptual capacities of animals he forthrightly denies mentality to the beasts:

> Animals have no knowledge of these sounds, but only a representation, without knowing whether what they apprehend is a sound or a color or something else; so one can say that they do not act so much as are acted upon, and that the objects make an impression upon their senses from which their action necessarily follows, as the wheels of a clock necessarily follow the weight or spring which drives them. (Mersenne 1636)

For Mersenne, then, the program inside animals is indeed an uninterpreted calculus, a syntax without a semantics [See Fodor: "Methodological Solipsism" *BBS* 3(1) 1980]. Searle, on the other hand, seems to believe that apes, monkeys, and dogs do "have mental states" because they "are made of similar stuff to ourselves" and have eyes, a nose, and skin. I fail to see how the datum supports the conclusion. One might have thought that some quite intricate reasoning and subtle experimentation would be required to justify the ascription of intentionality to chimpanzees (Marshall 1971; Woodruff & Premack 1979). That chimpanzees look quite like us is a rather weak fact on which to build such a momentous conclusion.

When Jacques de Vaucanson – the greatest of all AI theorists – had completed his artificial duck he showed it, in all its naked glory of wood, string, steel, and wire. However much his audience may have preferred a more cuddly creature, Vaucanson firmly resisted the temptation to clothe it:

> Perhaps some Ladies, or some People, who only like the Outside of Animals, had rather have seen the whole cover'd; that is the *Duck* with *Feathers*. But besides, that I have been desir'd to make every thing visible; I wou'd not be thought to impose upon the Spectators by any conceal'd or juggling Contrivance (Fryer & Marshall 1979).

For Vaucanson, the theory that he has embodied in the model duck is *the real thing*.

***Acknowledgment***

I thank Dr. J. Loew for his comments on earlier versions of this work.

*by* **Grover Maxwell**

*Center for Philosophy of Science, University of Minnesota, Minneapolis, Minn. 55455*

## Intentionality: Hardware, not software

It is a rare and pleasant privilege to comment on an article that surely is destined to become, almost immediately, a classic. But, alas, what comments are called for? Following *BBS* instructions, I'll resist the very strong temptation to explain how Searle makes exactly the right central points and supports them with exactly the right arguments; and I shall leave it to those who, for one reason or another, still disagree with his central contentions to call attention to a few possible weaknesses, perhaps even a mistake or two, in the treatment of some of his ancillary claims. What I shall try to do, is to examine, briefly – and therefore sketchily and inadequately – what seem to me to be some implications of his results for the overall mind-body problem.

Quite prudently, in view of the brevity of his paper, Searle leaves some central issues concerning mind-brain relations virtually untouched. In particular, his main thrust seems compatible with *interactionism*, with *epiphenomenalism*, and with at least some versions of the *identity thesis*. It does count, very heavily, against *eliminative materialism*, and , equally importantly, it reveals "functionalism" (or "functional materialism") as it is commonly held and interpreted (by, for example, Hilary Putnam and David Lewis) to be just another variety of eliminative materialism (protestations to the contrary notwithstanding). Searle correctly notes that functionalism of this kind (and strong AI, in general) is a kind of dualism. But it is not a mental-physical dualism; it is a form-content dualism, one, moreover, in which the form is the thing and content doesn't matter! [See Fodor: "Methodological Solipsism" *BBS* 3(1) 1980.]

Now I must admit that in order to find these implications in Searle's results I have read into them a little more than they contain explicity. Specifically, I have assumed that intentional states are genuinely mental in the what-is-it-like-to-be-a-bat? sense of "mental" (Nagel 1974) as well as, what I suppose is obvious, that eliminative materialism seeks to "eliminate" the genuinely mental in this sense. But it seems to me that it does not take much reading between the lines to see that Searle is sympathetic to my assumptions. For example, he does speak of "genuinely mental [systems]," and he says (in Searle 1979c) that he believes that "only beings capable of *conscious* states are capable of Intentional states" (my italics), although he says that he does not know how to demonstrate this. (How, indeed, could anyone demonstrate such a thing? How could one demonstrate that fire burns?)

The argument that Searle gives for the conclusion that only machines can think (can have intentional states) appears to have two suppressed premisses: (1) intentional states must always be causally produced, and (2) any causal network (with a certain amount of organization and completeness, or some such condition) is a machine. I accept for the purposes of this commentary his premises and his conclusion. Next I want to ask: what kind of hardware must a thinking machine incorporate? (By "thinking machine" I mean of course a machine that has genuinely mental thoughts: such a machine, I contend, will also have genuinely mental states or events instantiating sensations, emotions, and so on in all of their subjective, qualitative, conscious, experiential richness.) To continue this line of investigation, I want to employ an "event ontology," discarding substance metaphysics altogether. (Maxwell 1978, provides a sketch of some of the details and of the contentions that contemporary physics, quite independently of philosophy of mind, leads to such an ontology.) An event is (something like) the instancing of a property or the instancing (concrete realization) of a state. A causal network, then, consists entirely of a group of events and the causal links that interconnect them. A fortiori, our "machine" will consist entirely of events and causal connections. In other words, the hardware of this machine (or of any machine, for example, a refrigerator) consists of its constituent events and the machine consists of nothing else (except the causal linkages). Our thinking machine in the only form we know it today is always a brain (or if you prefer, an entire human or other animal body), which, as we have explained, is just a certain causal network of events. The mind-brain identity theory in the version that I defend says that some of the events in this network are (nothing but) genuinely mental events (instances of intentional states, or of pains, or the like). Epiphenomenalism says that the mental events "dangle" from the main net (the brain) by causal connections which are always one way (from brain to dangler). (Epiphenomenalism is, I believe, obviously, though contingently, false.) Interactionism says that there are causal connections in both directions but that the mental events are somehow in a different realm from the brain events. (How come a "different realm" – or whatever? Question: Is there a real difference between interactionism and identity theory in an event ontology?)

Assuming that Searle would accept the event ontology, if no more than for the sake of discussion, would he say that mental events, in general, and instances of intentional states, in particular, are parts of

the machine, or is it his position that they are just products of the machine? That is, would Searle be inclined to accept the identity thesis, or would he lean toward either epiphenomenalism or interactionism? For my money, in such a context, the identity theory seems by far the most plausible, elegant, and economical guess. To be sure, it must face serious and, as yet, completely unsolved problems, such as the "grain objection" (see, for example, Maxwell 1978), and emergence versus panpsychism (see, for example, Popper & Eccles 1977), but I believe that epiphenomenalism and interactionism face even more severe difficulties.

Before proceeding, I should emphasize that contemporary scientific knowledge not only leads us to an event ontology but also that it indicates the falsity of naive realism and "gently persuades" us to accept what I have (somewhat misleadingly, I fear) called "structural realism." According to this, virtually all of our knowledge of the physical world is knowledge of the structure (including space-time structure) of the causal networks that constitute it. (See, for example, Russell 1948 and Maxwell 1976). This holds with full force for knowledge about the brain (except for a very special kind of knowledge, to be discussed soon). We are, therefore, left ignorant as to what the intrinsic (nonstructural) properties of "matter" (or what contemporary physics leaves of it) are. In particular, if only we knew a lot more neurophysiology, we would know the structure of the (immense) causal network that constitutes the brain, but we would not know its content; that is, we still wouldn't know what any of its constituent events are. Identity theory goes a step further and speculates that some of these events just are (instances of) our intentional states, our sensations, our emotions, and so on, in all of their genuinely mentalistic richness, as they are known directly "by acquaintance." This is the "very special knowledge" mentioned above, and if identity theory is true, it is knowledge of what some (probably a very small subset) of the events that constitute the brain are. In this small subset of events we know intrinsic as well as structural properties.

Let us return to one of the questions posed by Searle: "could an artifact, a man-made machine, think?" The answer he gives is, I think, the best possible one, given our present state of unbounded ignorance in the neurosciences, but I'd like to elaborate a little. Since, I have claimed above, thoughts and other (genuinely) mental events are part of the hardware of "thinking machines," such hardware must somehow be got into any such machine we build. At present we have no inkling as to how this could be done. The best bet would seem to be, as Searle indicates, to "build" a machine (out of protoplasm) with neurons, dentrites, and axons like ours, and then to hope that, from this initial hardware, mental hardware would be mysteriously generated (would "emerge"). But this "best bet" seems to me extremely implausible. However, I do not conclude that construction of a thinking machine is (even contingently, much less logically) impossible. I conclude, rather, that we must learn a lot more about physics, neurophysiology, neuropsychology, psychophysiology, and so on, not just more details – but much more about the very foundations of our theroretical knowledge in these areas, before we can even speculate with much sense about building thinking machines. (I have argued in Maxwell 1978 that the foundations of contemporary physics are in such bad shape that we should hope for truly "revolutionary" changes in physical theory, that such changes may very well aid immensely in "solving the mind-brain problems," and that speculations in neurophysiology and perhaps even psychology may very well provide helpful hints for the physicist in his renovation of, say, the foundations of space-time theory.) Be all this as it may, Searle has shown the total futility of the strong AI route to genuine artificial intelligence.

*by* **E.W. Menzel, Jr.**

*Department of Psychology, State University of New York at Stony Brook, Stony Brook, N.Y. 11794*

## Is the pen mightier than the computer?

The area of artificial intelligence (AI) differs from that of natural intelligence in at least three respects. First, in AI one is perforce limited to the use of formalized behavioral data or "output" as a basis for making inferences about one's subjects. (The situation is no different, however, in the fields of history and archaeology.) Second, by convention, if nothing else, in AI one must ordinarily assume, until proved otherwise, that one's subject has no more mentality than a rock; whereas in the area of natural intelligence one can often get away with the opposite assumption, namely, that until proved otherwise, one's subject can be considered to be sentient. Third, in AI analysis is ordinarily limited to questions regarding the "structure" of intelligence, whereas a complete analysis of natural intelligence must also consider questions of function, development, and evolution.

In other respects, however, it seems to me that the problems of inferring mental capacities are very much the same in the two areas. And the whole purpose of the Turing test (or the many counterparts to that test which are the mainstay of comparative psychology) is to devise a clear set of rules for determining the status of subjects of any species, about whose possession of a given capacity we are uncertain. This is admittedly a game, and it cannot be freed of all possible arbitrary aspects any more than can, say, the field of law. Furthermore, unless everyone agrees to the rules of the game, there is no way to prove one's case for (or against) a given capacity with absolute and dead certainty.

As I see it, Searle quite simply refuses to play such games, at least according to the rules proposed by AI. He assigns himself the role of a judge who knows beforehand in most cases what the correct decision should be. And he does not, in my opinion, provide us with any decision rules for the remaining (and most interesting) undecided cases, other than rules of latter-day common sense (whose pitfalls and ambiguities are perhaps the major reason for devising objective tests that are based on performance rather than physical characteristics such as species, race, sex, and age.).

Let me be more specific. First of all, the discussion of "the brain" and "certain brain processes" is not only vague but seems to me to displace and complicate the problems it purports to solve. In saying this I do not imply that physiological data are irrelevant; I only say that their relevance is not made clear, and the problem of deciding where the brain leaves off and nonbrain begins is not as easy as it sounds. Indeed, I doubt that many neuroanatomists would even try to draw any sharp and unalterable line that demarcates exactly where in the animal kingdom "the brain" emerges from "the central nervous system"; and I suspect that some of them would ask, Why single out the brain as crucial to mind or intentionality? Why not the central nervous system or DNA or (to become more restrictive rather than liberal) the human brain or the Caucasian brain? Precisely analogous problems would arise in trying to specify for a single species such as man precisely how much intact brain, or what parts of it, or which of the "certain brain processes," must be taken into account and when one brain process leaves off and another one begins. Quite coincidentally, I would be curious as to what odds Searle would put on the likelihood that a neurophysiologist could distinguish between the brain processes of Searle during the course of his hypothetical experiment and the brain processes of a professor of Chinese. Also, I am curious as to what mental status he would assign to, say, an earthworm.

Second, it seems to me that, especially in the domain of psychology, there are always innumerable ways to skin a cat, and that these ways are not necessarily commensurate, especially when one is discussing two different species or cultures or eras. Thus, for example, I would be quite willing to concede that to "acquire the calculus" I would not require the intellectual power of Newton or of Leibnitz, who invented the calculus. But how would Searle propose to quantify the relative "causal powers" that are involved here, or how would he establish the relative similarity of the "effects"? The problem is especially difficult when Searle talks about subjects who have "zero understanding," for we possess no absolute scales or ratio scales in this domain, but only relativistic ones. In other words, we can assume by definition that a given subject may be taken as a criterion of "zero understanding," and assess the competence of other subjects by comparing them against this norm; but someone else is always free to invoke some other norm. Thus, for example, Searle uses himself as a standard of comparison and assumes he possesses zero understanding of Chinese. But what if I proposed that a better norm would be, say, a dog? Unless Searle's performance were no better than that of the dog, it seems to me that

the student of AI could argue that Searle's understanding must be greater than zero, and that his hypothetical experiment is therefore inconclusive; that is, the computer, which performs as he did, cannot necessarily be said to have zero understanding either.

In addition to these problems, Searle's hypothetical experiment is based on the assumption that AI would be proved "false" if it could be shown that even a single subject on a single occasion could conceivably pass a Turing test despite the fact that he possesses what may be assumed to be zero understanding. This, in my opinion, is a mistaken assumption. No student of AI would, to my knowledge, claim infallibility. His predictions would be at best probabilistic or statistical; and, even apart from problems such as cheating, some errors of classification are to be expected on the basis of "chance" alone. Turing, for example, predicted only that by the year 2000 computers will be able to fool an average interrogator on a Turing test, and be taken for a person, at least 30 times out of 100. In brief, I would agree with Searle if he had said that the position of strong AI is unprovable with dead certainty; but by his criteria no theory in empirical science is provable, and I therefore reject his claim that he has shown AI to be false.

Perhaps the central question raised by Searle's paper is, however, Where does mentality lie? Searle tells us that the intelligence of computers lies in our eyes alone. Einstein, however, used to say, "My pencil is more intelligent than I"; and this maxim seems to me to come at least as close to the truth as Searle's position. It is quite true that without a brain to guide it and interpret its output, the accomplishments of a pencil or a computer or of any of our other "tools of thought" would not be very impressive. But, speaking for myself, I confess I'd have to take the same dim view of my own accomplishments. In other words, I am quite sure that I could not even have "had" the thoughts expressed in the present commentary without the assistance of various means of externalizing and objectifying "them" and rendering them accessible not only for further examination but for their very formulation. I presume that there were some causal connections and correspondences between what is now on paper (or is it in the reader's eyes alone?) and what went on in my brain or mind; but it is an open question as to what these causal connections and correspondences were. Furthermore, it is only if one confuses present and past, and internal and external happenings with each other, and considers them a single "thing," that "thinking" or even the causal power behind thought can be allocated to a single "place" or entity. I grant that it is metaphorical if not ludicrous to give my pencil the credit or blame for the accomplishment of "the thoughts in this commentary." But it would be no less metaphorical and ludicrous – at least in science, as opposed to everyday life – to give the credit or blame to my brain as such. Whatever brain processes or mental processes were involved in the writing of this commentary, they have long since been terminated. In the not-too-distant future not even "I" as a body will exist any longer. Does this mean that the reader of the future will have no valid basis for estimating whether or not I was (or, as a literary figure, "am") any more intelligent than a rock? I am curious as to how Searle would answer this question. In particular, I would like to know whether he would ever infer from an artifact or document alone that its author had a brain or certain brain processes – and, if so, how this is different from making inferences about mentality from a subject's outputs alone.

*by* **Marvin Minsky**

***Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Mass. 02139***

## Decentralized minds

In this essay, Searle asserts without argument: "The study of the mind starts with such facts as that humans have beliefs, while thermostats, telephones, and adding machines don't. If you get a theory that denies this. . .the theory is false."

*No. The study of mind is not the study of belief; it is the attempt to* discover powerful concepts – be they old or new – that help explain why some machines or animals can do so many things others cannot. I will argue that traditional, everyday, precomputational concepts like believing and understanding are neither powerful nor robust enough for developing or discussing the subject.

In centuries past, biologists argued about machines and life much as Searle argues about programs and mind; one might have heard: "The study of biology begins with such facts as that humans have life, while locomotives and telegraphs don't. If you get a theory that denies this – the theory is false." Yet today a successful biological science is based on energetics and information processing; no notion of "alive" appears in the theory, nor do we miss it. The theory's power comes from replacing, for example, a notion of "self-reproduction" by more sophisticated notions about encoding, translation, and recombination – to explain the reproduction of sexual animals that do not, exactly, "reproduce."

Similarly in mind science, though prescientific idea germs like "believe," know," and "mean" are useful in daily life, they seem technically too coarse to support powerful theories; we need to supplant, rather than to support and explicate them. Real as "self" or "understand" may seem to us today, they are not (like milk and sugar) objective things our theories must accept and explain; they are only first steps toward better concepts. It would be inappropriate here to put forth my own ideas about how to proceed; instead consider a fantasy in which our successors recount our present plight: "The ancient concept of 'belief' proved inadequate until replaced by a continuum in which, it turned out, stones placed near zero, and termostats scored 0.52. The highest human score measured so far is 67.9. Because it is theoretically possible for something to be believed as intensely as 3600, we were chagrined to learn that men in fact believe so poorly. Nor, for that matter, are they very proficient (on an absolute scale) at intending. Still, they are comfortably separated from the thermostats." [Olivaw, R.D. (2063) Robotic reflections, *Phenomenological Science* 67:60.] A joke, of course; I doubt any such one-dimensional idea would help much. Understanding how parts of a program or mind can relate to things outside of it – or to other parts within – is complicated, not only because of the intricacies of hypothetical *intentions* and *meanings*, but because different parts of the mind do different things – both with regard to externals and to one another. This raises another issue: "In employing formulas like 'A believes that B means C,' our philosophical precursors became unwittingly entrapped in the 'single-agent fallacy' – the belief that inside each mind is a single believer (or meaner) who does the believing. It is strange how long this idea survived, even after Freud published his first clumsy portraits of our inconsistent and adversary mental constitutions. To be sure, that myth of 'self' is indispensable both for social purposes, and in each infant's struggle to make simplified models of his own mind's structure. But it has not otherwise been of much use in modern applied cognitive theory, such as our work to preserve, rearrange, and recombine those aspects of a brain's mind's parts that seem of value." [Byerly, S. (2080) New hope for the Dead, *Reader's Digest,* March 13.]

Searle talks of letting "the individual internalize all of these elements of the system" and then complains that "there isn't anything in the system that isn't in him." Just as our predecessors sought "life" in the study of biology, Searle still seeks "him" in the study of mind, and holds strong AI to be impotent to deal with the phenomenology of understanding. Because this is so subjective a topic, I feel it not inappropriate to introduce some phenomenology of my own. While reading about Searle's hypothetical person who incorporates into his mind – "without understanding" – the hypothetical "squiggle squoggle" process that appears to understand Chinese, I found my own experience to have some of the quality of a double exposure: "The text makes sense to some parts of my mind but, to other parts of my mind, it reads much as though it were itself written in Chinese. I understand its syntax, I can parse the sentences, and I can follow the technical deductions. But the terms and assumptions themselves – what the words like 'intend' and 'mean' intend and mean – escape me. They seem suspiciously like Searle's 'formally specified symbols' – because their principal meanings engage only certain older parts of my *mind that are not in harmonious, agreeable contact with just those* newer parts that seem better able to deal with such issues (precisely because they know how to exploit the new concepts of strong AI)."

Searle considers such internalizations – ones not fully integrated in the whole mind – to be counterexamples, or *reductiones ad absurdum* of some sort, setting programs somehow apart from minds. I see them

as illustrating the *usual* condition of normal minds, in which different fragments of structure interpret – and misinterpret – the fragments of activity they "see" in the others. There is absolutely no reason why programs, too, cannot contain such conflicting elements. To be sure, the excessive clarity of Searle's example saps its strength; the man's Chinese has no contact at all with his other knowledge – while even the parts of today's computer programs are scarcely ever jammed together in so simple a manner.

In the case of a mind so split into two parts that one merely executes some causal housekeeping for the other, I should suppose that each part – the Chinese rule computer and its host – would then have its own separate phenomenologies – perhaps along different time scales. No wonder, then, that the host can't "understand" Chinese very fluently – here I agree with Searle. But (for language, if not for walking or breathing) surely the most essential nuances of the experience of intending and understanding emerge, not from naked data bases of assertions and truth values, but from the interactions – the consonances and conflicts among different reactions within various partial selves and self-images.

What has this to do with Searle's argument? Well, I can see that if one regards intentionality as an all-or-none attribute, which each machine has or doesn't have, then Searle's idea – that intentionality emerges from some physical semantic principle – might seem plausible. But in may view this idea (of intentionality as a simple attribute) results from an oversimplification – a crude symbolization – of complex and subtle introspective activities. In short, when we construct our simplified models of our minds, we need terms to represent whole classes of such consonances and conflicts – and, I conjecture, this is why we create omnibus terms like "mean" and "intend." Then, those terms become reified.

It is possible that only a machine as decentralized yet interconnected as a human mind would have anything very like a human phenomenology. Yet even this supports no Searle-like thesis that mind's character depends on more than abstract information processing – on, for example, the "particular causal properties" of the substances of the brain in which those processes are embodied. And here I find Searle's arguments harder to follow. He criticizes *dualism,* yet complains about fictitious antagonists who suppose mind to be as substantial as sugar. He derides "residual operationalism" – yet he goes on to insist that, somehow, the chemistry of a brain can contribute to the quality or flavor of its mind with no observable effect on its behavior.

Strong AI enthusiasts do not maintain, as Searle suggests, that "what is specifically mental about the mind has *no* intrinsic connection with the actual properties of the brain." Instead, they make a much more discriminating scientific hypothesis: about *which* such causal properties are important in mindlike processes – namely *computation-supporting properties.* So, what Searle mistakenly sees as a difference in kind is really one of specific detail. The difference is important because what might appear to Searle as careless inattention to vital features is actually a deliberate – and probably beneficial – scientific strategy! For, as Putnam points out:

> *What is our intellectual form?* is the question, not what the matter is. Small effects may have to be explained in terms of the actual physics of the brain. But when are not even at the level of an *idealized* description of the functional organization of the brain, to talk about the importance of small perturbations seems decidedly premature.

Now, many strong AI enthusiasts go on the postulate that functional organization is the only such dependency, and it is this bold assumption that leads directly to the conclusion Searle seems to dislike so; that nonorganic machines may have the same kinds of experience as people do. That seems fine with me. I just can't see why Searle is so opposed to the idea that a *really* big pile of junk might have feelings like ours. He proposes no evidence whatever against it, he merely tries to portray it as absurd to imagine machines, with minds like ours – intentions and all – made from stones and paper instead of electrons and atoms. But I remain left to wonder how Searle, divested of despised dualism and operationalism, proposes to distinguish the authentic intentions of carbon compounds from their behaviorally identical but mentally counterfeit imitations.

I feel that I have dealt with the arguments about Chinese, and those about substantiality. Yet a feeling remains that there is something deeply wrong with all such discussions (as this one) of other minds; nothing ever seems to get settled in them. From the finest minds, on all sides, emerge thoughts and methods of low quality and little power. Surely this stems from a burden of traditional ideas inadequate to this tremendously difficult enterprise. Even our logic may be suspect. Thus, I could even agree with Searle that modern computational ideas are of little worth here – if, with him, I could agree to judge those ideas by their coherence and consistency with earlier constructions in philosophy. However, once one suspects that there are other bad apples in the logistic barrel, rigorous consistency becomes much too fragile a standard – and we must humbly turn to what evidence we can gather. So, beacuse this is still a formative period for our ideas about mind, I suggest that we must remain especially sensitive to the empirical power that each new idea may give us to pursue the subject further. And, as Searle nowhere denies, computationalism is the principal source of the new machines and programs that have produced for us the first imitations, however limited and shabby, of mindlike activity.

*by* **Thomas Natsoulas**

*Department of Psychology, University of California, Davis, Calif. 95616*

## The primary source of intentionality

I have shared Searle's belief: the level of description that computer programs exemplify is not one adequate to the explanation of mind. My remarks about this have appeared in discussions of perceptual theory that make little if any reference to computer programs per se (Natsoulas 1974; 1977; 1978a; 1978b; 1980). Just as Searle argues for the material basis of mind – "that actual human mental phenomena [depend] on actual physical-chemical properties of actual human brains" – I have argued that the particular concrete nature of perceptual awarenesses, as occurrences in a certain perceptual system, is essential to the references they make to objects, events, or situations in the stimulational environment.

In opposition to Gibson (for example, 1966; 1967; 1972), whose perceptual theory amounts to hypotheses concerning the pickup by perceptual systems of abstract entities called "informational invariants" from the stimulus flux, I have stated:

> Perceptual systems work *in their own modes* to ascribe the detected properties which are specified informationally to the actual physical environment around us. The informational invariants to which a perceptual system resonates are defined abstractly [by Gibson] such that the resonating process itself can exemplify them. But the resonance process is *not* itself abstract. And characterization of it at the level of informational invariants cannot suffice for the theory of perception. It is crucial to the theory of perception that informational invariants are resonated to *in concrete modes* that are characteristic of the organism as the kind of perceiving physical system that it is. (Natsoulas 1978b, p. 281)

The latter is crucial to perceptual theory, I have argued, if that theory is to explain the intentionality, or aboutness, of perceptual awarenesses [see also Ullman: "Against Direct perception" *BBS* 3(3) 1980, this issue].

And just as Searle summarily rejects the attempt to eliminate intentionality, saying that it does no good to "feign anesthesia," I argued as follows against Dennett's (1969; 1971; 1972) effort to treat intentionality as merely a heuristic overlay on the extensional theory of the nervous system and bodily motions.

> In knowing that we are knowing subjects, here is one thing that we know: that we are aware of objects in a way other than the "colorless" way in which we sometimes think of them. The experienced presence of objects makes it difficult if not impossible to claim that perceptions involve only the acquisition of information. . . It is this. . . kind of presence that makes perceptual aboutness something more than an "interpretation" or "heuristic overlay" to be dispensed with once a complete enough extensional account is at hand. The qualitative being thereness of objects and scenes. . . is about as easy to doubt as our own existence. (Natsoulas 1977, pp. 94–95; cf. Searle, 1979b, p. 261, on "presentational immediacy.")

However, I do not know in what perceptual aboutness consists. I have been making some suggestions and I believe, with Sperry (1969; 1970; 1976), that an objective description of subjective experience is possible in terms of brain function. Such a description should include that feature or those features of perceptual awareness that make it be of (or as if it is of, in the hallucinatory instance) an object, occurrence, or situation in the physical environment or in the perceiver's body outside the nervous system. If the description did not include this feature it would be incomplete, in my view, and in need of further development.

In another recent article on intentionality, Searle (1979a) has written of the unavoidability of "the intentional circle," arguing that any explanation of intentionality that we may come up with will presuppose an understanding of intentionality: "There is no analysis of intentionality into logically necessary and sufficient conditions of the form 'X is an intentional state S if and only if 'p, q, and r,' where 'p, q, and r' make no use of intentional notions" (p. 195). I take this to say that the intentionality of mental states is not reducible; but I don't think it is meant, by itself, to rule out the possibility that intentionality might be a property of certain brain processes. Searle could still take the view that it is one of their as yet unknown "ground floor" properties.

But Searle's careful characterization, in the target article, of the relation between brain and mental processes as causal, with mental processes consistently said to be *produced* by brain processes, gives a different impression. Of course brain processes produce other brain processes, but if he had meant to include mental processes among the latter, would he have written about only the *causal* properties of the brain in a discussion of the material basis of intentionality?

One is tempted to assume that Searle would advocate some form of interactionism with regard to the relation of mind to brain. I think that his analogy of mental processes to products of biological processes, such as sugar and milk, was intended to illuminate the causal basis of mental processes and not their nature. His statement that intentionality is "a biological phenomenon" was prefaced by "whatever else intentionality is" and was followed by a repetition of his earlier point concerning the material basis of mind (mental processes as produced by brain processes). And I feel quite certain that Searle would not equate mental processes with another of the brain's effects, namely behaviors (see Searle 1979b).

Though it may be tempting to construe Searle's position as a form of dualism, there remains the more probable alternative that he has simple chosen not to take, in these recent writings, a position on the ontological question. He has chosen to deal only with those elements that seem already clear to him as regards the problem of intentionality. However, his emphasis in the target article on intentionality's material basis would seem to be an indication that he is now inching toward a position on the ontological question and the view that this position *matters* to an understanding of intentionality.

I emphasize the latter because of what Searle has written on "the form of realization" of mental states in still another recent article on intentionality. In this article (Searle 1979c), he made the claim that the "traditional ontological problems about mental states are for the most part simply irrelevant to their Intentional features" (p. 81). It does not matter how a mental state is realized, Searle suggested, so long as in being realized it is realized so as to have those features. To know what an intentional state is requires only that we know its "representative content" and its "psychological mode."

But this would not tell us actually *what* the state is, only which one it is, or what kind it is. For example, I may know that the mental state that just occurred to me was a passing thought to the effect that it is raining in London at this moment. It is an it-is-raining-in-London-right-now kind of thought that just occurred to me. Knowing of this thought's occurrence and of that of many others, which is to know their representative contents and psychological modes, would not be to know what the thought is, what the mental occurrence is that is the passing thought.

Moreover, for a mental state or occurrence to have its intentional features, it must have a form of realization that gives it to them. Searle has stated: "It doesn't matter how an Intentional state is realized, as long as the realization is a realization of its Intentionality" (1979c, p. 81). The second part of this sentence is an admission that it *does* matter how it is realized. The explanation of intentionality remains incomplete in the absence of an account of its source and of its relation to that source. This point becomes vivid when considered in the context of another discussion contained in the same article.

In this part of the article, Searle gave some attention to what he called "the primary form of Intentionality," namely perception (cf. Searle 1979b, pp. 260–61). One's visual experience of a table was said to be a "presentation" of the table, as opposed to a representation of it. Still, such presentations are intrinsically intentional, for whenever they occur, the person thereby perceives or hallucinates a table. The visual experience if *of* a table, even though it is not a representation of a table, because it is satisfied by the physical presence of a table there where the table visually seems to be located.

Concluding this discussion, Searle added,

> To say that I have a visual experience whenever I perceive the table visually is to make an ontological claim, a claim about the form of realization of an Intentional state, in a way that to say all my beliefs have a representational content is not to make an ontological claim. (1979c, p. 91)

Since Searle would say that he, and the people reading his article, and animals, and so on, have visual experiences, the question he needs to answer, as the theorist of intentionality he is, is: what is the ontological claim he is making in doing so? Or, what is the "form of realization" of our visual experiences that Searle is claiming when he attributes them to us?

*by* **Roland Puccetti**

*Department of Philosophy, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5*

## The chess room: further demythologizing of strong AI

On the grounds he has staked out, which are considerable, Searle seems to me completely victorious. What I shall do here is to lift the sights of his argument and train them on a still larger, very tempting target.

Suppose we have an intelligent human from a chess-free culture, say Chad in Central Africa, and we introduce him to the chess room. There he confronts a computer console on which are displayed numbers 1–8 and letters R,N,B,K,Q, and P, plus the words WHITE and BLACK. He is told WHITE goes first, then BLACK, alternately, until the console lights go out. There is, of course, a binary machine representation of the chessboard that prevents illegal moves, but he need know nothing of that. He is instructed to identify himself with WHITE, hence to move first, and that the letter-number combination P-K4 is a good beginning. So he presses P-K4 and waits.

BLACK appears on the console, followed by three alternative letter-number combinations, P-K4, P-QB4, and P-K3. If this were a "depth-first" program, each of these replies would be searched two plies further and a static evaluation provided. Thus to BLACK's P-K4, WHITE could try either N-KB3 or B-B4. If N-KB3, BLACK's rejoinder N-QB3 or P-Q3 both yield an evaluation for WHITE of +1; whereas if B-B4, BLACK's reply of either B-B4 or N-KB3 produces an evaluation of, respectively, +0 and +3. Since our Chadian has been instructed to reject any letter-number combinations yielding an evaluation of less than +1, he will not pursue B-B4, but is prepared to follow N-KB3 unless a higher evaluation turns up. And in fact it does. The BLACK response P-QB4 allows N-KB3, and to that, BLACK's best countermoves P-Q3, P-K3, and N-QB3 produce evaluations of +7, +4, and +8. On the other hand, if this were a "breadth-first" program, in which all nodes (the point at which one branching move or half-move subdivides into many smaller branches in the game tree) at one level are examined prior to nodes at a deeper level, WHITE's continuations would proceed more statically; but again this does not matter to the Chadian in the chess room, who, in instantiating either kind of program, hasn't the foggiest notion what he is doing.

We must get perfectly clear what this implies. Both programs described here play chess (Frey 1977), and the latter with striking success in recent competition when run on a more powerful computer than before, a large scale Control Data Cyber 170 system (Frey 1977, Appendix). Yet there is not the slightest reason to believe either program *understands* chess play. Each performs "computational operations on purely formally specified elements," but so would the uncomprehending Chadian in our chess room, although of course much more slowly (we could probably use him only for postal chess, for this reason). Such operations, by themselves cannot, then, constitute understanding the game, no matter how intelligently played.

It is surprising that this has not been noticed before. For example, the authors of the most successful program to date (Slate & Atkin 1977) write that the evaluative function of CHESS 4.5 *understands* that it is bad to make a move that leaves one of one's own pieces attacked and undefended, it is good to make a move that forks two enemy pieces, and good to make a move that prevents a forking maneuver by the opponent (p. 114). Yet in a situation in which the same program is playing WHITE to move with just the WHITE king on KB5, the BLACK king on KR6, and BLACK's sole pawn advancing from KR4 to a possible queening, the initial evaluation of WHITE's six legal moves is as follows:

| Move | PREL score |
|---|---|
| K-K4 | −116 |
| K-B4 | −114 |
| K-N4 | −106 |
| K-K5 | −121 |
| K-K6 | −129 |
| K-B6 | −127 |

In other words, with a one-ply search the program gives a slightly greater preference to WHITE moving K-N4 because one of its evaluators encourages the king to be near the surviving enemy pawn, and N4 is as close as the WHITE king can legally get. This preliminary score does not differ much from that of the other moves since, as the authors admit, "the evaluation function does not understand that the pawn will be captured two half-moves later (p. 111)." It is only after a two-ply and then a three-ply iteration of K-N4 that the program finds that all possible replies are met. The authors candidly conclude:

> The whole 3-ply search here was completed in about 100 milliseconds. In a tournament the search would have gone out to perhaps 12 phy to get the same result, since the program lacks the sense to see that since White can force a position in which all material is gone, the game is necessarily drawn. (p. 113).

But then if CHESS 4.5 does not understand even this about chess, why say it "understands" forking maneuvers, and the like? All this can mean is that the program has built-in evaluators that discourage it from getting into forked positions and encourage it to look for ways to fork its opponent. That is not understanding, since as we saw, our Chadian in the chess room could laboriously achieve the same result on the console in blissful ignorance of chess boards, chess positions, or indeed how the game is played. Intelligent chess play is of course simulated this way, but chess understanding is not thereby duplicated.

Up until the middle of this century, chess-playing machines were automata with cleverly concealed human players inside them (Carroll 1975). We now have much more complex automata, and while the programs they run on are inside them, they do not have the intentionality towards the chess moves they make that midget humans had in the hoaxes of yesteryear. They simply know not what they do.

Searle quite unnecessarily mars his argument near the end of the target article by offering the observation, perhaps to disarm hard-nosed defenders of strong AI, that we humans are "thinking machines." But surely if he was right to invoke literal meaning against claims that, for example, thermostats have beliefs, he is wrong to say humans are machines of any kind. There were literally no machines on this planet 10,000 years ago, whereas the species *Homo sapiens* has existed here for at least 100,000 years, so it cannot be that men are machines.

*by* **Zenon W. Pylyshyn**[1]

*Center for Advanced Study in the Behavioral Sciences, Stanford, Calif. 94305.*

## The "causal power" of machines

***What kind of stuff can refer?*** Searle would have us believe that computers, qua formal symbol manipulators, necessarily lack the quality of intentionality, or the capacity to understand and to refer, because they have different "causal powers" from us. Although just what having different causal powers amounts to (other than not being capable of intentionality) is not spelled out, it appears at least that systems that are functionally identical need not have the same "causal powers." Thus the relation of equivalence with respect to causal powers is a refinement of the relation of equivalence with respect to function. What Searle wants to claim is that only systems that are equivalent to humans in this stronger sense can have intentionality. His thesis thus hangs on the assumption that intentionality is tied very closely to specific material properties – indeed, that it is literally *caused* by them. From that point of view it would be extremely unlikely that any system not made of protoplasm – or something essentially identical to protoplasm – can have intentionality. Thus if more and more of the cells in your brain were to be replaced by integrated circuit chips, programmed in such a way as to keep the input-output *function* of each unit identical to that of the unit being replaced, you would in all likelihood just keep right on speaking exactly as you are doing now except that you would eventually stop *meaning* anything by it. What we outside observers might take to be words would become for you just certain noises that circuits caused you to make.

Searle presents a variety of seductive metaphors and appeals to intuition in support of this rather astonishing view. For example, he asks: why should we find the view that intentionality is tied to detailed properties of the material composition of the system so surprising, when we so readily accept the parallel claim in the case of lactation? Surely it's obvious that only a system with certain causal powers can produce milk; but then why should the same not be true of the ability to refer? Why this example should strike Searle as even remotely relevant is not clear, however. The product of lactation is a *substance,* milk, whose essential defining properties are, naturally, physical and chemical ones (although nothing prevents the production of synthetic milk using a process that is materially very different from mammalian lactation). Is Searle then proposing that intentionality is a *substance* secreted by the brain, and that a possible test for intentionality might involve, say, titrating the brain tissue that realized some putative mental episodes?

Similarly, Searle says that it's obvious that merely having a program can't possibly be a sufficient condition for intentionality since you can implement that program on a Turing machine made out of "a roll of toilet paper and a pile of small stones." Such a machine would not have intentionality because such objects "are the wrong kind of stuff to have intentionality in the first place." But what is the right kind of stuff? Is it cell assemblies, individual neurons, protoplasm, protein molecules, atoms of carbon and hydrogen, elementary particles? Let Searle name the level, and it can be simulated perfectly well using "the wrong kind of stuff." Clearly it isn't the *stuff* that has the intentionality. Your brain cells don't refer any more than do the water pipes, bits of paper, computer operations, or the homunculus in the Chinese room examples. Searle presents no argument for the assumption that what makes the difference between being able to refer and not being able to refer – or to display any other capacity – is a "finer grained" property of the system than can be captured in a *functional* description. Furthermore, it's obvious from Searle's own argument that the nature of the stuff cannot be what is relevant, since the monolingual English speaker who has memeorized the formal rules is supposed to be an example of a system made of the *right* stuff and yet it allegedly still lacks the relevant intentionality.

Having said all this, however, one might still want to maintain that in some cases – perhaps in the case of Searle's example – it might be appropriate to say that *nothing* refers, or that the symbols are not being used in a way that refers to something. But if we wanted to deny

that these symbols referred, it would be appropriate to ask what licences us *ever* to say that a symbol refers. There are at least three different approaches to answering that question: Searle's view that it is the nature of the embodiment of the symbol (of the brain substance itself), the traditional functionalist view that it is the *functional role* that the symbol plays in the overall behavior of the system, and the view associated with philosophers like Kripke and Putnam, that it is in the nature of the causal connection that the symbol has with certain past events. The latter two are in fact compatible insofar as specifying the functional role of a symbol in the behavior of a system does not preclude specifying its causal interactions with an environment. It is noteworthy that Searle does not even consider the possibility that a purely formal computational model might constitute an essential part of an adequate theory, where the latter also contained an account of the system's transducers and an account of how the symbols came to acquire the role that they have in the functioning of the system.

***Functionalism and reference.*** The functionalist view is currently the dominant one in both AI and information-processing psychology. In the past, mentalism often assumed that reference was established by relations of similarity; an image referred to a horse if it *looked* sufficiently like a horse. Mediational behaviorism took it to be a simple causal remnant of perception: a brain event referred to a certain object if it shared some of the properties of brain events that occur when that object is perceived. But information-processing psychology has opted for a level of description that deals with the informational, or encoded, aspects of the environment's effects on the organism. On this view it has typically been assumed that what a symbol represents can be seen by examining how the symbol enters into relations with other symbols and with transducers. It is this position that Searle is quite specifically challenging. My own view is that although Searle is right in pointing out that some versions of the functionalist answer are in a certain sense incomplete, he is off the mark both in his diagnosis of where the problem lies and in his prognosis as to just how impoverished a view of mental functioning the cognitivist position will have to settle for (that is, his "weak AI").

The sense in which a functionalist answer might be incomplete is if it failed to take the further step of specifying what it was about the system that *warranted* the ascription of one particular semantic content to the functional states (or to the symbolic expressions that express that state) rather than some other logically possible content. A cognitive theory claims that the system behaves in a certain way *because* certain expressions represent certain things (that is, have a certain *semantic* interpretation). It is, furthermore, essential that it do so: otherwise we would not be able to subsume certain classes of regular behaviors in a single generalization of the sort "the system does X because the state S represents such and such" (for example, the person ran out of the building because he believed that *it was on fire).* (For a discussion of this issue, see Pylyshyn 1980b.) But the particular interpretation appears to be extrinsic to the theory inasmuch as the system would behave in exactly the same way without the interpretation. Thus Searle concludes that it is only *we,* the theorists, who take the expression to represent, say, that the building is on fire. The system doesn't take it to *represent* anything because it literally doesn't know what the expression refers to: only we theorists do. That being the case, the system can't be said to behave in a certain way *because* of what it represents. This is in contrast with the way in which *our* behavior is determined: we *do* behave in certain ways because of what our thoughts are about. And that, according to Searle, adds up to weak AI; that is, a functionalist account in which formal analogues "stands in" for, but themselves neither have, nor explain, mental contents.

The last few steps, however, are non sequiturs. The fact that it was we, the theorists, who provided the interpretation of the expressions doesn't by itself mean that such an interpretation is simply a matter of convenience, or that there is a sense in which the interpretation is ours rather than the system's. Of course it's logically possible that the interpretation is only in the mind of the theorist and that the system behaves the way it does for entirely different reasons. But even if that happened to be true, it wouldn't follow simply from the fact that the AI theorist was the one who came up with the interpretation. Much depends on his reasons for coming up with that interpretation. In any case, the question of whether the semantic interpretation resides in the head of the programmer or in the machine is the wrong question to ask. A more relevant question would be: what fixes the semantic interpretation of functional states, or what latitude does the theorist have in assigning a semantic interpretation to the states of the system?

When a computer is viewed as a self-contained device for processing formal symbols, we have a great deal of latitude in assigning semantic interpretations to states. Indeed, we routinely change our interpretation of the computer's functional states, sometimes viewing them as numbers, sometimes as alphabetic characters, sometimes as words or descriptions of a scene, and so on. Even where it is difficult to think of a coherent interpretation that is different from the one the programmer had in mind, such alternatives are always possible in principle. However, if we equip the machine with transducers and allow it to interact freely with both natural and linguistic environments, and if we endow it with the power to make (syntactically specified) inferences, it is anything but obvious what latitude, if any, the theorist (who knows how the transducers operate, and therefore knows what they respond to) would still have in assigning a coherent interpretation to the functional states in such a way as to capture psychologically relevant regularities in behavior.

***The role of intuitions.*** Suppose such connections between the system and the world as mentioned above (and possibly other considerations that no one has yet considered) uniquely constrained the possible interpretations that could be placed on representational states. Would this solve the problem of justifying the ascription of particular semantic contents to these states? Here I suspect that one would run into differences of opinion that may well be unresolvable, simply because they are grounded on different intuitions. For example there immediately arises the question of whether we possess a privileged interpretation of our own thoughts that must take precedence over such functional analyses. And if so, then there is the further question of whether being *conscious* is what provides the privileged access; and hence the question of what one is to do about the apparent necessity of positing unconscious mental processes. So far as I can see the *only* thing that recommends that particular view is the intuition that, whatever may be true of other creatures, I at least *know* what *my* thoughts refer to because I have direct experiential access to the referents of my thoughts. Even if we did have strong intuitions about such cases, there is good reason to believe that such intuitions should be considered as no more than secondary sources of constraint, whose validity should be judged by how well theoretical systems based on them perform. We cannot take as sacred anyone's intuitions about such things as whether another creature has intentionality – especially when such intuitions rest (as Searle's do, by his own admission) on knowing what the creature (or machine) is *made of* (for instance, Searle is prepared to admit that other creatures might have intentionality if "we can see that the beasts are made of similar stuff to ourselves"). Clearly, intuitions based on nothing but such anthropocentric chauvinism cannot form the foundation of a science of cognition [See "Cognition and Consciousness in Nonhuman Species" *BBS* 1(4) 1978].

A major problem in science – especially in a developing science like cognitive psychology – is to decide what sorts of phenomena "go together," in the sense that they will admit of a uniform set of explanatory principles. Information-processing theories have achieved some success in accounting for aspects of problem solving, language processing, perception, and so on, by deliberately glossing over the conscious-unconscious distinction; by gouping under common principles a wide range of rule-governed processes necessary to account for functioning, independent of whether or not people are aware of them. These theories have also placed to one side questions as to what constitute consciously experienced *qualia* or "raw feels" – dealing only with some of their reliable functional correlates (such as the belief that one is in pain, as opposed to the experience of the pain itself) – and they have to a large extent deliberately avoided the

question of what gives symbols their semantics. Because AI has chosen to carve up phenomena in this way, people like Searle are led to conclude that what is being done is weak AI – or the modelling of the abstract functional structure of the brain without regard for what its states represent. Yet there is no reason to think that this program does not in fact lead to strong AI in the end. There is no reason to doubt that at asymptote (for example, when and if a robot is built) the ascription of intentionality to programmed machines will be just as warranted as its ascription to people, and for reasons that have absolutely nothing to do with the issue of consciousness.

What is frequently neglected in discussions of intentionality is that we cannot state with any degree of precision what it is that entitles us to claim that *people* refer (though there are one or two general ideas, such as those discussed above), and therefore that arguments against the intentionality of computers typically reduce to "argument from ignorance." If we knew what it was that warranted our saying that people refer, we might also be in a position to claim that the ascription of semantic content to formal computational expressions – though it is in fact accomplished in practice by "inference to the best explanation" – was in the end warranted in exactly the same way. Humility, if nothing else, should prompt us to admit that there's a lot we don't know about how we ought to describe the capacities of future robots and other computing machines, even when we do know how their electronic circuits operate.

### *Note*

**1.** Current address: Department of Psychology, University of Western Ontario, London, Canada, N6A 5C2.

*by* **Howard Rachlin**

*Department of Psychology, State University of New York at Stony Brook, Stony Brook, N.Y. 11794*

## The behaviorist reply (Stony Brook)

It is easy to agree with the negative point Searle makes about mind and AI in this stimulating paper. What is difficult to accept is Searle's own conception of mind.

His negative point is that the mind can never be a computer program. Of course, that is what behaviorists have said all along ("residual behaviorism" in the minds of AI researchers notwithstanding). His positive point is that the mind is the same thing as the brain. But this is just as clearly false as the strong AI position that he criticizes.

Perhaps the behaviorist viewpoint can best be understood through two examples, one considered by Searle and one (although fairly obvious) not considered. The combination robot example is essentially a behavioral one. A robot behaves exactly like a man. Does it think? Searle says "If the robot looks and behaves sufficiently like us, then we would suppose, until *proven otherwise*, that it must have mental states like ours"(italics mine). Of couse we would. And let us be clear about what this robot would be required to do. It might answer questions about a story that it hears, but it should also laugh and cry in the right places; it should be able to tell when the story is over. If the story is a moral one the robot might change its subsequent behavior in situations similar to the ones the story describes. The robot might *ask* questions about the story itself, and the answers it receives might change its behavior later. The list of typically human behaviors in "response" to stories is endless. With a finite number of tests we can never be absolutely positive that the robot understood the story. But *proof otherwise* can only come from one place – the robot's subsequent behavior. That is, the robot may prove that it did not understand a story told to it at time-X by doing or saying something at a later time that a normal human would not do who heard a similar story under similar conditions. If it passes all our behavior tests we would say that, pending future *behavioral* evidence, the robot understood the story. And we would say this even if we were to open up the robot and find a man translating Chinese, a computer, a dog, a monkey, or a piece of stale cheese.

The appropriate test is to see whether the robot, upon hearing the story, behaves like a normal human being. How does a normal human being behave when told a story? That is a valid question – one in which behaviorists have been interested and one to which Searle and his fellow mentalists might also profitably devote their attention when they finish fantasizing about what goes on inside the head. The neural mythology that Searle suggests is no better than the computer-program mythology of the AI researchers.

Searle is willing to abandon the assumption of intentionality (in a robot) as soon as he discovers that a computer was running it after all. Here is a perfect example of how cognitive concepts can serve as a mask for ignorance. The robot is said to think until we find out how it works. Then it is said not to think. But suppose, contrary to anyone's expectations, all of the functional properties of the human brain were discovered. Then the "human robot" would be unmasked, and we might as well abandon the assumption of intentionality for humans too. It is only the behaviorist, it seems, who is willing to preserve terms such as *thought, intentionality,* and the like (as patterns of behavior). But there are no "mental states underlying. . . behavior" in the way that a skeleton underlies our bodily structure. The pattern of the behavior *is* the mental state. These patterns are results of internal and external factors in the present and in the past – not of a controlling mental state – even one identified with the brain.

That the identification of mind with brain will not hold up is obvious from the consideration of another example which I daresay will be brought up by other commentators – even AI researchers – so obvious is it. Let's call it "The Donovan's brain reply (Hollywood)." A brain is removed from a normal adult human. That brain is placed inside a computer console with the familiar input-output machinery – tape recorders, teletypewriters, and so on. The brain is connected to the machinery by a series of interface mechanisms that can stimulate all nerves that the body can normally stimulate and measure the state of all nerves that normally affect muscular movement. The brain, designed to interact with a body, will surely do no better (and probably a lot worse) at operating the interface equipment than a standard computer mechanism designed for such equipment. This "robot" meets Searle's criterion for a thinking machine – indeed it is an ideal thinking machine from his point of view. But it would be ridiculous to say that it could think. A machine that cannot behave like a human being cannot, by definition, think.

*by* **Martin Ringle**

*Computer Science Department, Vassar College, Poughkeepsie, N.Y. 12601*

## Mysticism as a philosophy of artificial intelligence

Searle identifies a weakness in AI methodology that is certainly worth investigating. He points out that by focusing attention at such a high level of cognitive analysis, AI ignores the foundational role that physical properties play in the determination of intentionality. The case may be stated thus: in human beings, the processing of cognized features of the world involves direct physical activity of neural structures and substructures as well as causal interactions between the nervous system and external physical phenomena. When we stipulate a "program" as an explanation (or, minimally, a description) of a cognitive process we abstract information-processing-type elements at some arbitrary level of resolution and we *presuppose* the constraints and contributions made at lower levels (for example, physical instantiation). AI goes wrong, according to Searle, by forgetting the force of this presupposition and thereby assuming that the computer implementation of the stipulated program will, by itself, display the intentional properties of the original human phenomenon.

AI doctrine, of course, holds that the lower-level properties are irrelevant to the character of the higher level cognitive processes – thus following the grand old tradition inaugurated by Turing (1964) and Putnam (1960).

If this is in fact the crux of the dispute between Searle and AI, then it is of relatively small philosophical interest. For it amounts to saying nothing more than that there may be important information processes occurring at the intraneuronal and subneuronal levels – a question that can only be decided empirically. If it turns out that such processes do not exist, then current approaches in AI are vindicated; if, on the other

hand, Searle's contention is correct, then AI must accommodate lower level processes in its cognitive models. Pragmatically, the simulation of subneuronal processes on a scale large enough to be experimentally significant might prove to be impossible (at least with currently envisioned technology). This is all too likely and would, if proven to be the case, spell methodological doom for AI as we now know it. Nevertheless, this would have little philosophical import since the inability to model the interface between complex subneural and interneuronal processes adequately would constitute a technical, not a theoretical, failure.

But Searle wants much more than this. He bases his denial of the adequacy of AI models on the belief that the physical properties of neuronal systems are such that they cannot *in principle* be simulated by a nonprotoplasmic computer system. This is where Searle takes refuge in what can only be termed mysticism.

Searle refers to the privileged properties of protoplasmic neuronal systems as "causal powers." I can discover at least two plausible interpretations of this term, but neither will satisfy Searle's argument. The first reading of "causal power" pertains to the direct linkage of the nervous system to physical phenomena of the external world. For example, when a human being processes visual images, the richness of the internal information results from direct physical interaction with the world. When a computer processes a scene, there need be no actual link between light phenomena in the world and an internal "representation" in the machine. Because the internal "representation" is the result of some stipulated program, one could (and often does in AI) input the "representation" by hand, that is, without any physical, visual apparatus. In such a case, the causal link between states of the world and internal states of the machine is merely stipulated. Going one step further, we can argue that without such a causal link, the internal states cannot be viewed as cognitive states since they lack any programmer-independent semantic content. AI workers might try to remedy the situation by introducing appropriate sensory transducers and effector mechanisms (such as "hands") into their systems, but I suspect that Searle could still press his point by arguing that the causal powers of such a system would still fail to mirror the *precise* causal powers of the human nervous system. The suppressed premise that Searle is trading on, however, is that nothing but a system that shared the physical properties of our systems would display precisely the same sort of causal links.

Yet if the causality with which Searle is concerned involves nothing more than direct connectivity between internal processes and sensorimotor states, it would seem that he is really talking about functional properties, not physical ones. He cannot make his case that a photo-electric cell is incapable of capturing the same sort of information as an organic rod or cone in a human retina unless he can specifically identify a (principled) deficiency of the former with respect to the latter. And this he does not do. We may sum up by saying that "causal powers," in this interpretation, *does* presuppose embodiment but that no particular physical makeup for a body is demanded. Connecting actual sensorimotor mechanisms to a perceptronlike internal processor should, therefore, satisfy causality requirements of this sort (by removing the stipulational character of the internal states).

Under the second interpretation, the term "causal powers" refers to the capacities of protoplasmic neurons to produce phenomenal states, such as felt sensations, pains, and the like. Here, Searle argues that things like automobiles and typewriters, because of their inorganic physical composition, are categorically incapable of causing felt sensations, and that this aspect of consciousness is crucial to intentionality.

There are two responses to this claim. First, arguing with Dennett, Schank, and others, we might say that Searle is mistaken in his view that intentionality necessarily requires felt sensations, that in fact the functional components of sensations are all that is required for a cognitive model. But, even if we accept Searle's account of intentionality, the claim stills seems to be untenable. The mere fact that mental phenomena such as felt sensations have been, historically speaking, confined to protoplasmic organisms in no way demonstrates that such phenomena could not arise in a nonprotoplasmic system. Such an assertion is on a par with a claim (made in antiquity) that only organic creatures such as birds or insects could fly. Searle explicitly and repeatedly announces that intentionality "is a biological phenomenon," but he never explains what sort of biological phenomenon it is, nor does he ever give us a reason to believe that there is a property or set of properties inherent in protoplasmic neural matter that could not, in principle, be replicated in an alternative physical substrate.

One can only conclude that the knowledge of the necessary connection between intentionality and protoplasmic embodiment is obtained through some sort of mystical revelation. This, of course, shouldn't be too troublesome to AI researchers who, after all, trade on mysticism as much as anyone in cognitive science does these days. And so it goes.

*by* **Richard Rorty**

*Department of Philosophy, Princeton University, Princeton, N.J. 08544*

## Searle and the special powers of the brain

Searle sets up the issues as would a fundamentalist Catholic defending transubstantiation. Suppose a demythologizing Tillichian theologian urges that we think of the Eucharist not in terms of substantial change, but rather in terms of significance in the lives of the faithful. The defender of orthodoxy will reply that the "natural supernatural distinction cannot be just in the eye of the beholder but must be intrinsic; otherwise it would be up to any beholder to treat anything as supernatural." (Compare Searle on the mental-nonmental distinction, p. 420) Theology, the orthodox say, starts with such facts as that the Catholic Eucharist is a supernatural event whereas a Unitarian minister handing around glasses of water isn't. Searle says that "the study of the mind starts with such facts as that humans have beliefs, while thermostats . . . and adding machines don't." In theology, the orthodox continue, one presupposes the reality and knowability of the supernatural. Searle says, "In 'cognitive sciences' one presupposes the reality and knowability of the mental." The orthodox think that the demythologizers are just changing the subject, since we know in advance that the distinction between the natural and the supernatural is a distinction between two sorts of entities having different special causal powers. We know that we can't interpret the Eucharist "functionally" in terms of its utility as an expression of our ultimate concern, for there could be such an expression without the elements actually changing their substantial form. Similarly, Searle knows in advance that a cognitive state "couldn't be just computational processes and their output because the computational processes and their output can exist without the cognitive state." Both the orthodox theologian and Searle criticize their opponents as "unashamedly behavioristic and operationalist."

Searle uses the example of being trained to answer enough questions in Chinese to pass a Turing test. The defender of transubstantiation would use the example of a layman disguised as a priest reciting the mass and fooling the parishioners. The initial reply to Searle's example is that if the training kept on for years and years so that Searle became able to answer *all* possible Chinese questions *in* Chinese, then he jolly well *would* understand Chinese. If you can fool all the people all of the time, behaviorists and operationalists say, it doesn't count as fooling any more. The initial reply to the orthodox theologian's example is that when Anglican priests perform Eucharist services what happens is functionally identical with what happens in Roman churches, despite the "defect of intention" in Anglican orders. When you get a body of worshippers as large as the Anglican communion to take the Eucharist without the necessary "special causal powers" having been present, that shows you that those powers weren't essential to the sacrament. Sufficiently *widely* accepted simulation *is* the real thing. The orthodox, however, will reply that a "consecrated" Anglican Host is no more the Body of Christ than a teddy bear is a bear, since the "special causal powers" are the essence of the matter. Similarly, Searle knows in advance that "only something that has the same causal powers as brains can have intentionality."

How does Searle know that? In the same way, presumably, as the orthodox theologian knows things. Searle knows what "mental" and "cognitive" and such terms mean, and so he knows that they can't be properly applied in the absence of brains – or, perhaps, in the absence

of something that is much like a brain in respect to "causal powers." How would we tell whether something was *sufficiently* like?, behaviorists and operationalists ask. Presumably they will get no answer until we discover enough about how the brain works to distinguish intentionality from mere simulations of intentionality. How might a neutral party judge the dispute between Anglicans and Roman Catholics about the validity of Anglican orders? Presumably he will have to wait until we discover more about God.

But perhaps the analogy is faulty: we moderns believe in brains but not in God. Still, even if we dismiss the theological analogue, we may have trouble knowing just *what* brain research is supposed to look for. We must discover *content* rather than merely form, Searle tells us, for mental states are "literally a product of the operation of the brain" and hence no conceivable program description (which merely gives a form, instantiable by many different sorts of hardware) will do. Behaviorists and operationalists, however, think the form-content and program-hardware distinctions merely heuristic, relative, and pragmatic. This is why they are, if not shocked, at least wary, when Searle claims "that actual human mental phenomena might be dependent on actual physical-chemical properties of actual human brains." If this claim is to be taken in a controversial sense, then it seems just a device for ensuring that the secret powers of the brain will move further and further back out of sight every time a new model of brain functioning is proposed. For Searle can tell us that every such model is merely a discovery of formal patterns, and that "mental content" has still escaped us. (He could buttress such a suggestion by citing Henri Bergson and Thomas Nagel on the ineffable inwardness of even the brute creation.) There is, after all, no great difference – as far as the form-content distinction goes – between building models for the behavior of humans and for that of their brains. Without further guidance about how to tell content when we finally encounter it, we may well feel that all research in the area

> is an arch wherethro'
> Gleams that untravell'd world whose margin fades
> For ever and for ever when I move. (Tennyson: *Ulysses*)

My criticisms of Searle should not, however, be taken as indicating sympathy with AI. In 1960 Putnam remarked that the mind-program analogy did not show that we can use computers to help philosophers solve the mind-body problem, but that there wasn't any mind-body problem for philosophers to solve. The last twenty years' worth of work in AI have reinforced Putnam's claim. Nor, alas, have they done anything to help the neurophysiologists – something they actually *might*, for all we could have known, have done. Perhaps it was worth it to see whether programming computers could produce some useful models of the brain, if not of "thought" or "the mind." Perhaps, however, the money spent playing Turing games with expensive computers should have been used to subsidize relatively cheap philosophers like Searle and me. By now we might have worked out exactly which kinds of operationalism and behaviorism to be ashamed of and which not. Granted that some early dogmatic forms of these doctrines were a bit gross, Peirce was right in saying something like them has *got* to be true if we are to shrug off arguments about transubstantiation. If Searle's present pre-Wittgensteinian attitude gains currency, the good work of Ryle and Putnam will be undone and "the mental" will regain its numinous Cartesian glow. This will boomerang in favor of AI. "Cognitive scientists" will insist that only lots more simulation and money will shed empirical light upon these deep "philosophical" mysteries. Surely Searle doesn't want *that*.

***by* Roger C. Schank**

***Department of Computer Science, Yale University, New Haven, Conn, 06520***

## Understanding Searle

What is understanding? What is consciousness? What is meaning? What does it mean to think? These, of course, are philosopher's questions. They are the bread and butter of philosophy. But what of the role of such questions in AI? Shouldn't AI researchers be equally concerned with such questions? I believe the answer to be yes and no.

According to the distinction between weak and strong AI, I would have to place myself in the weak AI camp with a will to move to the strong side. In a footnote, Searle mentions that he is not saying that I am necessarily committed to the two "AI claims" he cites. He states that claims that computers can understand stories or that programs can explain understanding in humans are unsupported by my work.

He is certainly right in that statement. No program we have written can be said to truly understand yet. Because of that, no program we have written "explains the human ability to understand."

I agree with Searle on this for two reasons. First, we are by no means finished with building understanding machines. Our programs are at this stage partial and imcomplete. They cannot be said to be truly understanding. Because of this they cannot be anything more than partial explanations of human abilities.

Of course, I realize that Searle is making a larger claim than this. He means that our programs never will be able to understand or explain human abilities. On the latter claim he is clearly quite wrong. Our programs have provided successful embodiments of theories that were later tested on human subjects. All experimental work in psychology to date has shown, for example, that our notion of a script (Schank & Abelson 1977) is very much an explanation of human abilities (see Nelson & Gruendel 1978; Gruendel 1980; Smith, Adams, & Schorr 1978; Bower, Black, & Turner 1979; Graesser et al. 1979; Anderson 1980).

All of the above papers are reports of experiments on human subjects that support the notion of a script. Of course, Searle can hedge here and say that it was our theories rather than our programs that explained human abilities in that instance. In that case, I can only attempt to explain carefully my view of what AI is all about. We cannot have theories apart from our computer implementations of those theories. The range of the phenomena to be explained is too broad and detailed to be covered by a theory written in English. We can only know if our theories of understanding are plausible if they can work by being tested on a machine.

Searle is left with objecting to psychological experiments themselves as adequate tests of theories of human abilities. Does he regard psychology as irrelevant? The evidence suggests that he does, although he is not so explicit on this point. This brings me back to his first argument. "Can a machine understand?" Or, to it put another way, can a process model of understanding tell us something about understanding? This question applies whether the target of attack is AI or psychology.

To answer this question I will attempt to draw an analogy. Try to explain what "life" is. We can give various biological explanations of life. But, in the end, I ask, what is the essence of life? What is it that distinguishes a dead body that is physically intact from a live body? Yes, of course, the processes are ongoing in the live one and not going (or "dead") in the dead one. But how to start them up again? The jolt of electricity from Dr. Frankenstein? What is the "starter"? What makes life?

Biologists can give various process explanations of life, but in the end that elusive "starter of life" remains unclear. And so it is with understanding and consciousness.

We attribute understanding, consciousness, and life to others on the grounds that we ourselves have these commodities. We really don't know if anyone else "understands," "thinks," or even is "alive." We assume it on the rather unscientific basis that since we are all these things, others must be also.

We cannot give scientific explanations for any of these phenomena. Surely the answers, formulated in chemical terms, should not satisfy Searle. I find it hard to believe that what philosophers have been after for centuries were chemical explanations for the phenomena that pervade our lives.

Yet, that is the position that Searle forces himself into. Because, apart from chemical explanation, what is left? We need explanations in human terms, in terms of the entities that we meet and deal with in our daily lives, that will satisfy our need to know about these things.

Now I will return to my analogy. Can we get at the process explanation of "life"? Yes, of course, we could build a model that functioned "as if it were alive," a robot. Would it be alive?

The same argument can be made with respect to consciousness and understanding. We could build programs that functioned as if they understood or had free conscious thought. Would they be conscious? Would they really understand?

I view these questions somewhat differently from most of my AI colleagues. I do not attribute beliefs to thermostats, car engines, or computers. My answers to the above questions are tentative no's. A robot is not alive. Our story-understanding systems do not understand in the sense of the term that means true empathy of feeling and expression.

Can we ever hope to get our programs to "understand" at that level? Can we ever create "life"? Those are, after all, empirical questions.

In the end, my objection to Searle's remarks can be formulated this way. Does the brain understand? Certainly we humans understand, but does that lump of matter we call our brain understand? All that is going on there is so many chemical reactions and electrical impulses, just so many Chinese symbols.

Understanding means finding the system behind the Chinese symbols, whether written for brains or for computers. The person who wrote the rules for Searle to use to put out the correct Chinese symbols at the appropriate time – now that was a linguist worth hiring. What rules did he write? The linguist who wrote the rules "understood" in the deep sense how the Chinese language works. And, the rules he wrote embodied that understanding.

Searle wants to call into question the enterprise of AI, but in the end, even he must appreciate that the rules for manipulating Chinese symbols would be a great achievement. To write them would require a great understanding of the nature of language. Such rules would satisfy many of the questions of philosophy, linguistics, psychology, and AI.

Does Searle, who is using those rules, understand? No. Does the hardware configuration of the computer understand? No. Does the hardware configuration of the brain understand? No.

Who understands then? Why, the person who wrote the rules of course. And who is he? He is what is called an AI researcher.

*by* **Aaron Sloman and Monica Croucher**

*School of Social Sciences, University of Sussex, Brighton BN1 9QN, England*

## How to turn an information processor into an understander

Searle's delightfully clear and provocative essay contains a subtle mistake, which is also often made by AI researchers who use familiar mentalistic language to describe their programs. The mistake is a failure to distinguish form from function.

That some mechanism or process has properties that would, in a suitable context, enable it to perform some function, does not imply that it already performs that function. For a process to be understanding, or thinking, or whatever, it is not enough that it replicate some of the structure of the processes of understanding, thinking, and so on. It must also fulfil the functions of those processes. This requires it to be causally linked to a larger system in which other states and processes exist. Searle is therefore right to stress causal powers. However, it is not the causal powers of brain cells that we need to consider, but the causal powers of computational processes. The reason the processes he describes do not amount to understanding is not that they are not produced by things with the right causal powers, but that *they* do not have the right causal powers, since they are not integrated with the right sort of total system.

That certain operations on symbols occurring in a computer, or even in another person's mind, happen to be isomorphic with certain formal operations in your mind does not entail that they serve the same function in the political economy of your mind. When you read a sentence, a complex, mostly unconscious, process of syntactic and semantic analysis occurs, along with various inferences, alterations to your long-term memory, perhaps changes in your current plans, or even in your likes, dislikes, or emotional state. Someone else reading the sentence will at most share a subset of these processes. Even if there is a subset of formal symbolic manipulations common to all those who hear the sentence, the existence of those formal processes will not, in isolation, constitute understanding the sentence. Understanding can occur only in a context in which the process has the opportunity to interact with such things as beliefs, motives, perceptions, inferences, and decisions – because it is embedded in an appropriate way in an appropriate overall system.

This may look like what Searle calls "The robot reply" attributed to Yale. However, it is not enough to say that the processes must occur in some physical system which it causes to move about, make noises, and so on. We claim that it doesn't even have to be a physical system: the properties of the larger system required for intentionality are *computational* not *physical*. (This, unlike Searle's position, explains why it makes sense to ordinary folk to attribute mental states to disembodied souls, angels, and the like, though not to thermostats.)

What sort of larger system is required? This is not easy to answer. There is the beginning of an exploration of the issues in chapters 6 and 10 of Sloman (1978) and in Sloman (1979). (See also Dennett 1978.) One of the central problems is to specify the conditions under which it could be correct to describe a computational system, whether embodied in a human brain or not, as possessing its own desires, preferences, tastes, and other motives. The conjecture we are currently exploring is that such motives are typically instantiated in symbolic representations of states of affairs, events, processes, or selection criteria, which play a role in controlling the operations of the system, including operations that change the contents of the store of motives, as happens when we manage (often with difficulty) to change our own likes and dislikes, or when an intention is abandoned because it is found to conflict with a principle. More generally, motives will control the allocation of resources, including the direction of attention in perceptual processes, the creation of goals and subgoals, the kind of information that is processed and stored for future use, and the inferences that are made, as well as controlling external actions if the system is connected to a set of 'motors' (such as muscles) sensitive to signals transmitted during the execution of plans and strategies. Some motives will be capable of interacting with beliefs to produce the complex disturbances characteristic of emotional states, such as fear, anger, embarrassment, shame, and disgust. A precondition for the system to have its own desires and purposes is that its motives should evolve as a result of a feedback process during a lengthy sequence of experiences, in which beliefs, skills (programs), sets of concepts, and the like also develop. This, in turn requires the system of motives to have a multilevel structure, which we shall not attempt to analyse further here.

This account looks circular because it uses mentalistic terminology, but our claim, and this is a claim not considered by Searle, is that further elaboration of these ideas can lead to a purely formal specification of the computational architecture of the required system. Fragments can already be found in existing operating systems (driven in part by priorities and interrupts), and in AI programs that interpret images, build and debug programs, and make and execute plans. But not existing system comes anywhere near combining all the intricacies required before the familiar mental processes can occur. Some of the forms are already there, but not yet the functions.

Searle's thought experiment, in which he performs uncomprehending operations involving Chinese symbols does not involve operations linked into an appropriate system in the appropriate way. The news, in Chinese, that his house is on fire will not send him scurring home, even though in some way he operates correctly with the symbols. But, equally, none of the so-called understanding programs produced so far is linked to an appropriate larger system of beliefs and decision. Thus, as far as the ordinary meanings of the words are concerned, it is incorrect to say that any existing AI programs understand, believe, learn, perceive, or solve problems. Of course, it might be argued (though not by us) that they already have the potential to be so linked – they have a form that is adequate for the function in question. If this were so, they might perhaps be used as extensions of people – for example, as aids for the deaf or blind or the mentally handicapped, and they could then be part of an understanding system.

It could be argued that mentalistic language should be extended to

encompass all systems with the *potential* for being suitably linked into a complete mind. That is, it could be argued that the meanings of words like "understand," "perceive," "intend," "believe" should have their functional preconditions altered, as if we were to start calling things screwdrivers or speed controllers if they happened to have the appropriate structure to perform the functions, whether or not they were ever used or even intended to be used with the characteristic functions of screwdrivers and speed controllers. The justification for extending the usage of intentional and other mental language in this way would be the discovery that some aspects of the larger architecture (such as the presence of subgoal mechanisms or inference mechanisms) seem to be required within such isolated subsystems to enable them to satisfy even the formal preconditions. However, our case against Searle does not depend on altering meanings of familiar words.

Is it necessary that a mental system be capable of controlling the operations of a physcial body or that it be linked to physical sensors capable of receiving information about the physical environment? This is close to the question whether a totally paralysed, deaf, blind, person without any functioning sense organs might nevertheless be conscious, with thoughts, hopes, and fears. (Notice that this is not too different from the state normal people enter temporarily each night.) We would argue that there is no reason (apart from unsupportable behaviourist considerations) to deny that this is a logical possiblility. However, if the individual had never interacted with the external world in the normal way, then he could not think of President Carter, Paris, the battle of Hastings, or even his own body: at best his thoughts and experiences would refer to similar nonexistent entities in an imaginary world. This is because successful reference presupposes causal relationships which would not hold in the case of our disconnected mind.

It might be thought that we have missed the point of Searle's argument since whatever the computational architecture we finally posit for a mind, connected or disconnected, he will always be able to repeat his thought experiment to show that a purely formal symbol-manipulating system with that structure would not necessarily have motives, beliefs, or percepts. For he could execute all the programs himself (at least in principle) without having any of the alleged desires, beliefs, perceptions, emotions, or whatever.

At this point the "other minds" argument takes on a curious twist. Searle is assuming that he is a final authority on such questions as whether what is going on in his mental activities includes seeing (or appearing to see) pink elephants, thinking about Pythagoras's theorem, being afraid of being burnt at the stake, or understanding Chinese sentences. In other words, he assumes, without argument, that it is impossible for another mind to be based on his mental processes without his knowing. However, we claim (compare the discussion of consciousness in Sloman 1978, chapter 10) that if he really does faithfully execute all the programs, providing suitable time sharing between parallel subsystems where necessary, then a collection of mental processes will occur of whose nature he will be ignorant, if all he thinks he is doing is manipulating meaningless symbols. He will have no more basis for denying the existence of such mental processes than he would have if presented with a detailed account of the low-level internal workings of another person's mind, which he can only understand in terms of electrical and chemical processes, or perhaps sequences of abstract patterns embedded in such processes.

If the instructions Searle is executing require him to use information about things *he* perceives in the environment as a basis for selecting some of the formal operatons, then it would even be possible for the "passenger" to acquire information about Searle (by making inferences from Searle's behaviour and from what other people say about him) without Searle ever realising what is going on. Perhaps this is not too unlike what happens in some cases of multiple personalities?

**by William E. Smythe**

*Department of Psychology, University of Toronto, Toronto, Ontario, Canada M5S 1A1*

## Simulation games

Extensive use of intentional idioms is now common in discussions of the capabilities and functioning of AI systems. Often these descriptions are to be taken no more substantively than in much of ordinary programming where one might say, for example, that a statistical regression program "wants" to minimize the sum of squared deviations or "believes" it has found a best-fitting function when it has done so. In other cases, the intentional account is meant to be taken more literally. This practice requires at least some commitment to the claim that intentional states can be achieved in a machine just in virtue of its performing certain computations. Searle's article serves as a cogent and timely indicator of some of the pitfalls that attend such a claim.

If certain AI systems are to possess intentionality, while other computational systems do not, then it ought to be in virtue of some set of purely computational principles. However, as Searle points out, no such principles have yet been forthcoming from AI. Moreover, there is reason to believe that they never will be. A sketch of one sort of argument is as follows: intentional states are, by definition, "directed at" objects and states of affairs in the world. Hence the first requirement for any theory about them would be to specify the relation between the states and the world they are "about." However it is precisely this relation that is not part of the computational account of mental states (cf. Foder 1980). A computational system can be interfaced with an external environment in any way a human user may choose. There is no dependence of this relation on any ontogenetic or phylogenetic history of interaction with the environment. In fact, the relation between system and environment can be anything at all without affecting the computations performed on symbols that purportedly refer to it. This fact casts considerable doubt on whether any purely computational theory of intentionality is possible.

Searle attempts to establish an even stronger conclusion: his argument is that the computational realization of intentional states is, in fact, impossible on a priori grounds. The argument is based on a "simulation game" – a kind of dual of Turing's imitation game – in which man mimics computer. In the simulation game, a human agent instantiates a computer program by performing purely syntactic operations on meaningless symbols. The point of the demonstration is that merely following rules for the performance of such operations is not sufficient for manifesting the right sort of intentionality. In particular, a given set of rules could create an effective mimicking of some intelligent activity without bringing the rule-following agent any closer to having intentional states pertaining to the domain in question.

One difficulty with this argument is that it does not distinguish between two fundamentally different ways of instantiating a computer program or other explicit prodecure in a physical system. One way is to imbed the program in a system that is already capable of interpreting and following rules. This requires that the procedure be expressed in a "language" that the imbedding system can already "understand." A second way is to instantiate the program *directly* by realizing its "rules" as primitive hardware operations. In this case a rule is followed, not by "interpreting" it, but by just running off whatever procedure the rule denotes. Searle' simulation game is germane to the first kind of instantiation but not the second. Following rules in natural language (as the simulation game requires) involves the mediation of other intentional states and so is necessarily an instance of indirect instantiation. To mimic a direct instantiation of a program faithfully, on the other hand, the relevant primitives would have to be realized nonmediately in one' own activity. If such mimicry were possible, it would be done only at the cost of being unable to report on the system's lack of intentional states, if in fact it had none.

The distinction between directly and indirectly instantiating computational procedures is important because both kinds of processes are required to specify a computational system completely. The first comprises its architecture or set of primitives, and the second comprises the algorithms the system can apply (Newell 1973; 1980). Hence Searle's argument is a challenge to strong AI when that view is put forward in terms of the capabilities of programs, but not when it is framed (as, for example, by Pylyshyn 1980a) in terms of computational systems. The claim that the latter cannot have intentional states must therefore proceed along different lines. The approach considered earlier, for example, called attention to the arbitrary relation between computational symbol and referent. Elsewhere the argument has been put forward in more detail that it is an overly restrictive notion of *symbol*

that creates the most serious difficulties for the computational theory (Kolers & Smythe 1979; Smythe 1979). The notion of an independent token subject to only formal syntactic manipulation is neither a sufficient characterization of what a symbol is, nor well motivated in the domain of human cognition. Sound though this argument is, it is not the sort of secure conclusion that Searle's simulation game tries to demonstrate.

However, the simulation game does shed some light on another issue. Why is it that the belief is so pervasive that AI systems are truly constitutive of mental events? One answer is that many people seem to be playing a different version of the simulation game from the one that Searle recommends. The symbols of most AI and cognitive simulation systems are rarely the kind of meaningless tokens that Searle's simulation game requires. Rather, they are often externalized in forms that carry a good deal of surplus meaning to the user, over and above their procedural identity in the system itself, as pictorial and linguistic inscriptions, for example. This sort of realization of the symbols can lead to serious theoretical problems. For example, systems like that of Kosslyn and Shwartz (1977) give the appearance of operating on mental images largely because their internal representations "look" like images when displayed on a cathode ray tube. It is unclear that the system could be said to manipulate images in any other sense. There is a similar problem with language understanding systems. The semantics of such systems is often assessed by means of an informal procedure that Hayes (1977, p. 559) calls "pretent-it's-English." That is, misleading conclusions about the capabilities of these systems can result from the superficial resemblance of their internal representations to statements in natural language. An important virute of Searle's argument is that it specifies how to play the simulation game correctly. The procedural realization of the symbols is all that should matter in a computational theory; their external appearance ought to be irrelevant.

The game, played this way, may not firmly establish that computational systems lack intentionality. However, it at least undermines one powerful tacit motivation for supposing they have it.

*by* **Donald O. Walter**

*Brain Research Institute and Department of Psychiatry, University of California, Los Angeles, Calif. 90024*

## The thermostat and the philosophy professor

***Searle:*** The man certainly doesn't understand Chinese, and neither do the water pipes, and if we are tempted to adopt what I think is the absurd view that somehow the *conjunction* of man *and* water pipes understands. . .

***Walter:*** The bimetallic strip by itself certainly doesn't keep the temperature within limits, and neither does the furnace by itself, and if we are tempted to adopt the view that somehow a *system* of bimetallic strip and furnace will keep the temperature within limits – or (paraphrasing Hanson 1969; or others), Searle's left retina does not see, nor does his right, nor either (or both) optic nerve(s); we can even imagine a "disconnection syndrome" in which Searle's optic cortex no longer connects with the rest of his brain, and so conclude that his optic cortex doesn't see, either. If we then conclude that because no part sees, therefore he cannot see, are we showing consistency, or are we failing to see something about our own concepts?

***Searle:*** No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down . . . Why on earth would anyone suppose that a computer simulation of understanding actually understood anything?

***Walter:*** No one supposes that a novelist's description of a five-alarm fire will burn the neighborhood down; why would anyone suppose that a novelist writing about understanding actually understood it?

***Searle:*** If we knew independently how to account for its behavior without such assumptions we would not attribute intentionality to it, especially if we knew it had a formal program.

***Hofstadter*** (1979, p. 601): There is a related "Theorem" about progress in AI: once some mental function is programmed, people soon cease to consider it as an essential ingredient of "real thinking." The ineluctable core of intelligence is always that next thing which hasn't yet been programmed.

***Walter:*** Searle seems to be certain that a program is formal (though he plays, to his own advantage, on the ambiguity between "adequately definable through form or shape" and "completely definable through nothing but form or shape"), whereas "intentionality," "causal powers" and "actual properties" are radically different things that are unarguably present in any (normal? waking?) human brain, and possibly in the quaint brains of "Martians" (if they were "alive," at least in the sense that we did not understand what went on inside them). These radically different things are also not definable in terms of their form but of their content. He asserts this repeatedly, without making anything explicit of this vital alternative. I think it is up to Searle to establish communication with the readers of this journal, which he has not done in his target article. Let us hope that in his Response he will make the mentioned but undescribed alternative more nearly explicit to us.

*by* **Robert Wilensky**

*Department of Electrical Engineering and Computer Science, University of California, Berkeley, Calif. 94720*

## Computers, cognition and philosophy

Searle's arguments on the feasibility of computer understanding contain several simple but fatal logical flaws. I can deal only with the most important difficulties here. However, it is the general thrust of Searle's remarks rather than the technical flaws in his arguments that motivates this commentary. Searle' paper suggests that even the best simulation of intelligent behavior would explain nothing about cognition, and he argues in support of this claim. Since I would like to claim that computer simulation can yield important insights into the nature of human cognitive processes, it is important to show why Searle's arguments do not threaten this enterprise.

My main objection to Searle's argument he has termed the "Berkeley systems reply." The position states that the man-in-the-room scenario presents no problem to a strong AI-er who claims that understanding is a property of an information-processing *system*. The man in the room with the ledger, functioning in the manner prescribed by the cognitive theorist who instructed his behavior, constitutes one such system. The man functioning in his normal everyday manner is another system. The "ordinary man" system may not understand Chinese, but this says nothing about the capabilities of the "man-in-the-room" system, which must therefore remain at least a candidate for consideration as an understander in view of its language-processing capabilities.

Searle's response to this argument is to have the man internalize the "man-in-the-room" system by keeping all the rules and computations in his head. He now encompasses the whole system. Searle argues that if the man "doesn't understand, then there is no way the system could understand because the system is just a part of him."

However, this is just plain wrong. Lots of systems (in fact, most interesting systems) are embodied in other systems of weaker capabilities. For example, the hardware of a computer may not be able to multiply polynomials, or sort lists, or process natural language, although programs written for those computers can; individual neurons probably don't have much – if any – understanding capability, although the systems they constitute may understand quite a bit.

The difficulty in comprehending the systems position in the case of Searle's paradox is in being able to see the person as two separate systems. The following elaborations may be useful. Suppose we decided to resolve the issue once and for all simply by asking the person involved whether he understands Chinese. We hand the person a piece of paper with Chinese characters that mean (loosely translated) "Do you understand Chinese?" If the man-in-the-room system were to respond, by making the appropriate symbol manipulations, it would return a strip of paper with the message: "Of course I understand Chinese! What do you think I've been doing? Are you joking?" A heated dialogue then transpires, after which we apologize to the man-in-the-room system for our rude innuendos. Immediately

thereafter, we approach the man himself (that is, we ask him to stop playing with the pieces of paper and talk to us directly) and ask him if he happens to know Chinese. He will of course deny such knowledge.

Searle's mistake of identifying the experiences of one system with those of its implementing system is one philosophers often make when referring to AI systems. For example, Searle says that the English subsystem knows that "hamburgers" refer to hamburgers, but that the Chinese subsystem knows only about formal symbols. But it is really the homunculus who is conscious of symbol manipulation, and has no idea what higher level task he is engaged in. The parasitic system is involved in this higher level task, and has no knowledge at all that he is implemented via symbol manipulation, anymore than we are aware of how our own cognitive processes are implemented.

What's unusual about this situation is not that one system is embedded in a weak one, but that the implementing system is so much more powerful than it need be. That is, the homunculus is a full-fledged understander, operating at a small percentage of its capacity to push around some symbols. If we replace the man by a device that is capable of performing only these opertaions, the temptation to view the systems as identical greatly diminishes.

It is important to point out, contrary to Searle's claim, that the systems position itself does not constitute a strong AI claim. It simply shows that *if* it is possible that a system other than a person functioning in the standard manner can understand, then the man-in-the-room argument is not at all problematic. If we deny this possibility to begin with, then the delicate man-in-the-room argument is unnecessary – a computer program is something other than a person functioning normally, and by assumption would not be capable of understanding.

Searle also puts forth an argument about simulation in general. He states that since a simulation of a storm won't leave us wet, why should we assume that a simulation of understanding should understand? Well, the reason is that while simulations don't necessarily preserve all the properties of what they simulate, they don't necessarliy violate particular properties either. I could simulate a storm in the lab by spraying water through a hose. If I'm interested in studying particular properties, I don't have to abandon simulations; I merely have to be careful about which properties the simulation I construct is likely to preserve.

So it all boils down to the question, what sort of thing is understanding? If it is an inherently physical thing, like fire or rain or digestion, then preserving the logical properties of understanding will in fact not preserve the essential nature of the phenomenon, and a computer simulation will not understand. If, on the other hand, understanding is essentially a logical or symbolic type of activity, then preserving its logical properties would be sufficient to have understanding, and a computer simulation will literally understand.

Searle's claim is that the term "understanding" refers to a physical phenomenon, much in the same way that the term "photosynthesis" does. His argument here is strictly an appeal to our intuitions about the meaning of this term. My own intuitions simply do not involve the causal properties of biological organisms (although they do involve their logical and behavioral properties). It seems to me that this must be true for most people, as most people could be fooled into thinking that a computer simulation really understands, but a simulation of photosynthesis would not fool anyone into thinking it had actually created sugar from water and carbon dioxide.

A major theme in Searle's paper is that intentionality is really at the bottom of the problem. Computers fail to meet the critieria of true understanding because they just don't have intentional states, with all that entails. This, according to Searle, is in fact what boggles one's intuitions in the man-in-the-room example.

However, it seems to me that Searle's argument has nothing to do with intentionality at all. What causes difficultly in attributing intentional states to machines is the fact that most of these states have a *subjective* nature as well. If this is the case, then Searle's man-in-the-room example could be used to simulate a person having some nonintentional but subjective state, and still have its desired effect. This is precisely what happens. For example, suppose we simulated someone undergoing undirected anxiety. It's hard to believe that anything – the man doing the simulation or the system he implements – is actually experiencing undirected anxiety, even though this is not an intentional state.

Futhermore, the experience of discomfort seems proportional to subjectivity, but independent of intentionality. It doesn't bother my intuitions much to hear that a computer can understand or know something; that it is believing something is a little harder to swallow, and that it has love, hate, rage, pain, and anxiety are much worse. Notice that the subjectivity seems to increase in each case, but the intentionality remains the same. The point is that Searle's argument has nothing to do with intentionality per se, and sheds no light on the nature of intentional states or on the kinds of mechanisms capable of having them.

I'd like to sum up by saying one last word on Searle's man-in-the-room experiment, as this forms the basis for most of his subsequent arguments. Woody Allen in *Without Feathers* describes a mythical beast called the Great Roe. The Great Roe has the head of a lion, and the body of a lion – but not the same lion. Searle's *Gedankenexperiment* is really a Great Roe – the head of an understander and the body of an understander, but not the same understander. Herein lies the difficulty.

# Author's Response

*by* **John Searle**

*Department of Philosophy, University of California, Berkeley, Calif. 94720*

## Intrinsic intentionality

I am pleased at the amount of interest my target article has aroused and grateful that such a high percentage of the commentaries are thoughtful and forcefully argued. In this response I am going to attempt to answer every major criticism directed at my argument. To do that, however, I need to make fully explicit some of the points that were implicit in the target article, as these points involve recurring themes in the commentaries.

**Strong AI.** One of the virtues of the commentaries is that they make clear the extreme character of the strong AI thesis. The thesis implies that of all known types of specifically biological processes, from mitosis and meiosis to photosynthesis, digestion, lactation, and the secretion of auxin, one and only one type is completely independent of the biochemistry of its origins, and that one is cognition. The reason it is independent is that cognition consists entirely of computational processes, and since those processes are purely formal, any substance whatever that is capable of instantiating the formalism is capable of cognition. Brains just happen to be one of the indefinite number of different types of computers capable of cognition, but computers made of water pipes, toilet paper and stones, electric wires – anything solid and enduring enough to carry the right program – will necessarily have thoughts, feelings, and the rest of the forms of intentionality, because that is all that intentionality consists in: instantiating the right programs. The point of strong AI is not that if we built a computer big enough or complex enought to carry the actual programs that brains presumably instantiate we would get intentionality as a byproduct (contra **Dennett**), but rather that there isn't anything to intentionality other than instantiating the right program.

Now I find the thesis of strong AI incredible in every sense of the word. But it is not enough to find a thesis incredible, one has to have an argument, and I offer an argument that is very simple: instantiating a program could not be constitutive of intentionality, because it would be possible for an agent to instantiate the program and still not have the right kind of

intentionality. That is the point of the Chinese room example. Much of what follows will concern the force of that argument.

**Intuitions.** Several commentators (**Block, Dennett, Pylyshyn, Marshall**) claim that the argument is just based on intuitions of mine, and that such intuitions, things we feel ourselves inclined to say, could never prove the sort of thing I am trying to prove (Block), or that equally valid contrary intuitions can be generated (Dennett), and that the history of human knowledge is full of the refutation of such intuitions as that the earth is flat or that the table is solid, so intuitions here are of no force.

But consider. When I now say that I at this moment do not understand Chinese, that claim does not merely record an intuition of mine, something I find myself inclined to say. It is a plain fact about me that I don't understand Chinese. Furthermore, in a situation in which I am given a set of rules for manipulating uninterpreted Chinese symbols, rules that allow no possibility of attaching any semantic content to these Chinese symbols, it is still a fact about me that I do not understand Chinese. Indeed, it is the very same fact as before. But, **Wilensky** suggests, suppose that among those rules for manipulating symbols are some that are Chinese for "Do you understand Chinese?," and in response to these I hand back the Chinese symbols for "Of course I understand Chinese." Does that show, as Wilensky implies, that there is a subsystem in me that understands Chinese? As long as there is no semantic content attaching to these symbols, the fact remains that there is no understanding.

The form of **Block**'s argument about intuition is that since there are allegedly empirical data to show that thinking is just formal symbol manipulation, we could not refute the thesis with untutored intuitions. One might as well try to refute the view that the earth is round by appealing to our intuition that it is flat. Now Block concedes that it is not a matter of intuition but a plain fact that our brains are "the seat" of our intentionality. I want to add that it is equally a plain fact that I don't understand Chinese. My paper is an attempt to explore the logical consequences of these and other such plain facts. Intuitions in his deprecatory sense have nothing to do with the argument. One consequence is that the formal symbol manipulations could not be constitutive of thinking. Block never comes to grips with the arguments for this consequence. He simply laments the feebleness of our intuitions.

**Dennett** thinks that he can generate counterintuitions. Suppose, in the "robot reply," that the robot is my very own body. What then? Wouldn't I understand Chinese then? Well, the trouble is that the case, as he gives it to us, is underdescribed, because we are never told what is going on in the mind of the agent. (Remember, in these discussions, always insist on the first person point of view. The first step in the operationalist sleight of hand occurs when we try to figure out how we would *know* what it would be like for others.) If we describe Dennett's case sufficiently explicitly it is not hard to see what the facts would be. Suppose that the program contains such instructions as the following: when somebody holds up the squiggle-squiggle sign, pass him the salt. With such instructions it wouldn't take one long to figure out that "squiggle squiggle" probably means pass the salt. But now the agent is starting to learn Chinese from following the program. But this "intuition" doesn't run counter to the facts I was pointing out, for what the agent is doing in such a case is attaching a semantic content to a formal symbol and thus taking a step toward language comprehension. It would be equally possible to describe a case in such a way that it was impossible to attach any semantic content, even though my own body was in question, and in such a case it would be impossible for me to learn Chinese from following the program. Dennett's examples do not generate counterintuitions, they are simply so inadequately described that we can't tell from his description what the facts would be.

At one point **Dennett** and I really do have contrary intuitions. He says "*I* understand English my brain doesn't." I think on the contrary that when I understand English; it is my brain that is doing the work. I find nothing at all odd about saying that my brain understands English, or indeed about saying that my brain is conscious. I find his claim as implausible as insisting, "I digest pizza; my stomach and digestive tract don't."

**Marshall** suggests that the claim that thermostats don't have beliefs is just as refutable by subsequent scientific discovery as the claim that tables are solid. But notice the difference. In the case of tables we discovered previously unknown facts about the microstructure of apparently solid objects. In the case of thermostats the relevant facts are all quite well known already. Of course such facts as that thermostats don't have beliefs and that I don't speak Chinese are, like all empirical facts, subject to disconfirmation. We might for example discover that, contrary to my deepest beliefs, I am a competent speaker of Mandarin. But think how we would establish such a thing. At a minimum we would have to establish that, quite unconsciously, I know the meanings of a large number of Chinese expressions; and to establish that thermostats had beliefs, in exactly the same sense that I do, we would have to establish, for example, that by some miracle thermostats had nervous systems capable of supporting mental states, and so on. In sum, though in some sense intuition figures in any argument, you will mistake the nature of the present dispute entirely if you think it is a matter of my intuitions against someone else's, or that some set of contrary intuitions has equal validity. The claim that I don't speak Chinese and that my thermostat lacks beliefs aren't just things that I somehow find myself mysteriously inclined to say.

Finally, in response to **Dennett** (and also **Pylyshyn** ), I do not, of course, think that intentionality is a fluid. Nor does anything I say commit me to that view. I think, on the contrary, that intentional states, processes, and events are precisely that: states, processes, and events. The point is that they are both caused by and realized in the structure of the brain. Dennett assures me that such a view runs counter to "the prevailing winds of doctrine." So much the worse for the prevailing winds.

**Intrinsic intentionality and observer-relative ascriptions of intentionality.** Why then do people feel inclined to say that, in some sense at least, thermostats have beliefs? I think that in order to understand what is going on when people make such claims we need to distinguish carefully between cases of what I will call *intrinsic intentionality,* which are cases of actual mental states, and what I will call *observer-relative ascriptions of intentionality,* which are ways that people have of speaking about entities figuring in our activities but lacking intrinsic intentionality. We can illustrate this distinction with examples that are quite uncontroversial. If I say that I am hungry or that Carter believes he can win the election, the form of intentionality in question is intrinsic. I am discussing, truly or falsely, certain psychological facts about me and Carter. But if I say the word "Carter" refers to the present president, or the sentence "Es regnet" means it's raining, I am not ascribing any mental states to the word "Carter" or the sentence "Es regnet." These are ascriptions of intentionality made to entities that lack any mental states, but in which the ascription is a manner of speaking about the intentionality of the observers. It is a way of saying that people *use* the name Carter to refer, or that when people say literally "Es regnet" they *mean* it's raining.

Observer-relative ascriptions of intentionality are always

dependent on the intrinsic intentionality of the observers. There are not two kinds of intentional mental states; there is only one kind, those that have intrinsic intentionality; but there are ascriptions of intentionality in which the ascription does not ascribe intrinsic intentionality to the subject of the ascription. Now I believe that a great deal of the present dispute rests on a failure to appreciate this distinction. When **McCarthy** stoutly maintains that thermostats have beliefs, he is confusing observer-relative ascriptions of intentionality with ascriptions of intrinsic intentionality. To see this point, ask yourself why we make these attributions to thermostats and the like at all. It is not because we suppose they have a mental life very much like our own; on the contrary, we know that they have no mental life at all. Rather, it is because we have designed them (our intentionality) to serve certain of our purposes (more of our intentionality), to perform the sort of functions that we perform on the basis of our intentionality. I believe it is equally clear that our ascription of intentionality to cars, computers, and adding machines is observer relative.

Functionalism, by the way, is an entire system erected on the failure to see this distinction. Functional attributions are always observer relative. There is no such thing as an intrinsic function, in the way that there are intrinsic intentional states.

**Natural kinds.** This distinction between intrinsic intentionality and observer-relative ascriptions of intentionality might seem less important if we could, as several commentators (**Minsky, Block, Marshall**) suggest, assimilate intrinsic intentionality to some larger natural kind that would subsume both existing mental phenomena and other natural phenomena under a more general explanatory apparatus. Minsky says that "prescientific idea germs like 'believe' " have no place in the mind science of the future (presumably "mind" will also have no place in the "mind science" of the future). But even if this is true, it is really quite irrelevant to my argument, which is addressed to the mind science of the present. Even if, as Minsky suggests, we eventually come to talk of our present beliefs as if they were on a continuum with things that are not intentional states at all, this does not alter the fact that we do have intrinsic beliefs and computers and thermostats do not. That is, even if some future science comes up with a category that supersedes belief and thus enables us to place thermostats and people on a single continuum, this would not alter the fact that under our present concept of belief, people literally have beliefs and thermostats don't. Nor would it refute my diagnosis of the mistake of attributing intrinsic mental states to thermostats as based on a confusion between intrinsic intentionality and observer-relative ascriptions of intentionality.

**Minsky** further points out that our own mental operations are often split into parts that are not fully integrated by any "self" and only some of which carry on interpretation. And, he asks, if that is how it is in our own minds, why not in computers as well? The reply is that even if there are parts of our mental processes where processing takes place without any intentional content, there still have to be other parts that attach semantic content to syntactic elements if there is to be any understanding to all. The point of the Chinese room example is that the formal symbol manipulations never by themselves carry any semantic content, and thus instantiating a computer program is not by itself sufficient for understanding.

**How the brain works.** Several commentators take me to task because I don't explain how the brain works to produce intentionality, and at least two (**Dennett** and **Fodor**) object to my claim that where intentionality is concerned – as opposed to the conditions of satisfaction of the intentionality – what matters are the internal and not the external causes. Well I don't know *how* the brain produces mental phenomena, and apparently no one else does either, but *that* it produces mental phenomena and that the internal operations of the brain are causally sufficient for the phenomena is fairly evident from what we do know.

Consider the following case, in which we do know a little about how the brain works. From where I am seated, I can see a tree. Light reflected from the tree in the form of photons strikes my optical apparatus. This sets up a series of sequences of neural firings. Some of these neurons in the visual cortex are in fact remarkably specialized to respond to certain sorts of visual stimuli. When the whole set of sequences occurs, it causes a visual experience, and the visual experience has intentionality. It is a conscious mental event with an intentional content; that is, its conditions of satisfaction are internal to it. Now I could be having exactly that visual experience even if there were no tree there, provided only that something was going on in my brain sufficient to produce the experience. In such a case I would not *see* the tree but would be having a hallucination. In such a case, therefore, the intentionality is a matter of the *internal* causes; whether the intentionality is satisfied, that is, whether I actually see a tree as opposed to having a hallucination of the tree, is a matter of the *external* causes as well. If I were a brain in a vat I could have exactly the same mental states I have now; it is just that most of them would be false or otherwise unsatisfied. Now this simple example of visual experience is designed to make clear what I have in mind when I say that the operation of the brain is causally sufficient for intentionality, and that it is the operation of the brain and not the impact of the outside world that matters for the content of our intentional states, in at least one important sense of "content."

Some of the commentators seem to suppose that I take the causal powers of the brain by themselves to be an argument against strong AI. But that is a misunderstanding. It is an empirical question whether any given machine has causal powers equivalent to the brain. My argument against strong AI is that instantiating a program is not enough to guarantee that it has those causal powers.

**Wait till next year.** Many authors (**Block, Sloman & Croucher, Dennett, Lycan, Bridgeman, Schank**) claim that Schank's program is just not good enough but that newer and better programs will defeat my objection. I think this misses the point of the objection. My objection would hold against any program at all, qua formal computer program. Nor does it help the argument to add the causal theory of reference, for even if the formal tokens in the program have some causal connection to their alleged referents in the real world, as long as the agent has no way of knowing that, it adds no intentionality whatever to the formal tokens. Suppose, for example, that the symbol for egg foo yung in the Chinese room is actually causally connected to egg foo yung. Still, the man in the room has no way of knowing that. For him, it remains an uninterpreted formal symbol, with no semantic content whatever. I will return to this last point in the discussion of specific authors, especially **Fodor.**

**Seriatim.** I now turn, with the usual apologies for brevity, from these more general considerations to a series of specific arguments.

**Haugeland** has an argument that is genuinely original. Suppose a Chinese speaker has her neurons coated with a thin coating that prevents neuron firing. Suppose "Searle's demon" fills the gap by stimulating the neurons as if they had been fired. Then she will understand Chinese even though none of her neurons has the right causal powers; the demon has them, and he understands only English.

My objection is only to the last sentence. Her neurons still have the right causal powers; they just need some help from the demon. More generally if the stimulation of the causes is

at a low enough level to *reproduce* the causes and not merely *describe* them, the "simulation" will reproduce the effects. If what the demon does is reproduce the right causal phenomena, he will have reproduced the intentionality, which constitutes the effects of that phenomena. And it does not, for example, show that my brain lacks the capacity for consciousness if someone has to wake me up in the morning by massaging my head.

**Haugeland's** distinction between original and derivative intentionality is somewhat like mine between intrinsic intentionality and observer-relative ascriptions of intentionality. But he is mistaken in thinking that the only distinction is that original intentionality is "sufficiently rich" in its "semantic activity": the semantic activity in question is still observer-relative and hence not sufficient for intentionality. My car engine is, in his observer-relative sense, semantically active in all sorts of "rich" ways, but it has no intentionality. A human infant is semantically rather inactive, but it still has intentionality.

**Rorty** sets up an argument concerning transubstantiation that is formally parallel to mine concerning intrinsic and observer-relative attributions of intentionality. Since the premises of the transubstantiation argument are presumably false, the parallel is supposed to be an objection to my argument. But the parallel is totally irrelevant. Any valid argument whatever from true premises to true conclusions has exact formal analogues from false premises to false conclusions. Parallel to the familiar "Socrates is mortal" argument we have "Socrates is a dog. All dogs have three heads. Therefore Socrates has three heads." The possibility of such formal parallels does nothing to weaken the original arguments. To show that the parallel was insightful Rorty would have to show that my premises are as unfounded as the doctrine of transubstantiation. But what are my premises? They are such things as that people have mental states such as beliefs, desires, and visual experiences, that they also have brains, and that their mental states are causally the products of the operation of their brains. Rorty says nothing whatever to show that these propositions are false, and I frankly can't suppose that he doubts their truth. Would he like evidence for these three? He concludes by lamenting that if my views gain currency the "good work" of his favorite behaviorist and functionalist authors will be "undone." This is not a prospect I find at all distressing, since implicit in my whole account is the view that people really do have mental states, and to say so is not just to ascribe to them tendencies to behave, or to adopt a certain kind of stance toward them, or to suggest functional explanations of their behaviours. This does not give the mental a "numinous Cartesian glow," it just implies that mental processes are as real as any other biological processes.

**McCarthy** and **Wilensky** both endorse the "systems reply." The major addition made by Wilensky is to suppose that we ask the Chinese subsystem whether it speaks Chinese and it answers yes. I have already suggested that this adds no plausibility whatever to the claim that there is any Chinese understanding going on in the system. Both Wilensky and McCarthy fail to answer the three objections I made to the systems reply.

1. The Chinese subsystem still attaches no semantic content whatever to the formal tokens. The English subsystem knows that "hamburger" means hamburger. The Chinese subsystem knows only that squiggle squiggle is followed by squoggle squoggle.

2. The systems reply is totally unmotivated. Its only motivation is the Turing test, and to appeal to that is precisely to beg the question by assuming what is in dispute.

3. The systems reply has the consequence that all sorts of systematic input-output relations (for example, digestion) would have to count as understanding, since they warrant as much observer-relative ascription of intentionality as does the Chinese subsystem. (And it is, by the way, no answer to this point to appeal to the cognitive impenetrability of digestion, in Pylyshyn's [1980a] sense, since digestion is congnitively penetrable: the content of my beliefs can upset my digestion.)

**Wilensky** seems to think that it is an objection that other sorts of mental states besides intentional ones could have been made the subject of the argument. But I quite agree. I could have made the argument about pains, tickles, and anxiety, but these are (a) less interesting to me and (b) less discussed in the AI literature. I prefer to attack strong AI on what its proponents take to be their strongest ground.

**Pylyshyn** misstates my argument. I offer no a priori proof that a system of integrated circuit chips couldn't have intentionality. That is, as I say repeatedly, an empirical question. What I do argue is that in order to produce intentionality the system would have to duplicate the causal powers of the brain and that simply instantiating a formal program would not be sufficient for that. Pylyshyn offers no answer to the arguments I give for these conclusions.

Since **Pylyshyn** is not the only one who has this misunderstanding, it is perhaps worth emphasizing just what is at stake. The position of strong AI is that anything with the right program would have to have the relevant intentionality. The circuit chips in his example would necessarily have intentionality, and it wouldn't matter if they were circuit chips or water pipes or paper clips, provided they instantiated the program. Now I argue at some length that they couldn't have intentionality solely in virtue of instantiating the program. Once you see that the program doesn't necessarily add intentionality to a system, it then becomes an empirical question which kinds of systems really do have intentionality, and the condition necessary for that is that they must have causal powers equivalent to those of the brain. I think it is evident that all sorts of substances in the world, like water pipes and toilet paper, are going to lack those powers, but that is an empirical claim on my part. On my account it is a testable empirical claim whether in repairing a damaged brain we could duplicate the electrochemical basis of intentionality using some other substance, say silicon. On the position of strong AI there cannot be any empirical questions about the electrochemical bases necessary for intentionality since any substance whatever is sufficient for intentionality if it has the right program. I am simply trying to lay bare for all to see the full preposterousness of that view.

I believe that **Pylyshyn** also misunderstands the distinction between intrinsic and observer-relative ascriptions of intentionality. The relevant question is not how much latitude the observer has in making observer-relative ascriptions, but whether there is any intrinsic intentionality in the system to which the ascriptions could correspond.

**Schank** and I would appear to be in agreement on many issues, but there is at least one small misunderstanding. He thinks I want "to call into question the enterprise of AI." That is not true. I am all in favor of weak AI, at least as a research program. I entirely agree that if someone could write a program that would give the right input and output for Chinese stories it would be a "great achievement" requiring a "great understanding of the nature of language." I am not even sure it can be done. My point is that instantiating the program is not constitutive of understanding.

**Abelson,** like **Schank,** points out that it is no mean feat to program computers that can simulate story understanding. But, to repeat, that is an achievement of what I call weak AI, and I would enthusiastically applaud it. He mars this valid point by insisting that since our own understanding of most things, arithmetic for example, is very imperfect, "we might well be humble and give the computer the benefit of the doubt when and if it performs as well as we do." I am afraid that neither this nor his other points meets my arguments to

show that, humble as we would wish to be, there is no reason to suppose that instantiating a formal program in the way a computer does is any reason *at all* for ascribing intentionality to it.

**Fodor** agrees with my central thesis that instantiating a program is not a sufficient condition of intentionality. He thinks, however, that if we got the right causal links between the formal symbols and things in the world that would be sufficient. Now there is an obvious objection to this variant of the robot reply that I have made several times: the same thought experiment as before applies to this case. That is, no matter what outside causal impacts there are on the formal tokens, these are not by themselves sufficient to give the tokens any intentional content. No matter what caused the tokens, the agent still doesn't understand Chinese. Let the egg foo yung symbol be causally connected to egg foo yung in any way you like, that connection by itself will never enable the agent to interpret the symbol as meaning egg foo young. To do that he would have to have, for example, some *awareness* of the causal relation between the symbol and the referent; but now we are no longer explaining intentionality in terms of symbols and causes but in terms of symbols, causes, and intentionality, and we have abandoned both strong AI and the robot reply. Fodor's only answer to this is to say that it shows we haven't yet got the right kind of causal linkage. But what is the right kind, since the above argument applies to any kind? He says he can't tell us, but it is there all the same. Well I can tell him what it is: it is any form of causation sufficient to produce intentional content in the agent, sufficient to produce, for example, a visual experience, or a memory, or a belief, or a semantic interpretation of some word.

**Fodor's** variant of the robot reply is therefore confronted with a dilemma. If the causal linkages are just matters of fact about the relations between the symbols and the outside world, they will never by themselves give any interpretation to the symbols; they will carry by themselves no intentional content. If, on the other hand, the causal impact is sufficient to produce intentionality in the agent, it can only be because there is something more to the system than the *fact* of the causal impact and the *symbol*, namely the intentional content that the impact produces in the agent. Either the man in the room doesn't learn the meaning of the symbol from the causal impact, in which case the causal impact adds nothing to the interpretation, or the causal impact teaches him the meaning of the word, in which case the cause is relevant only because it produces a form of intentionality that is something in addition to itself and the symbol. In neither case is symbol, or cause and symbol, constitutive of intentionality.

This is not the place to discuss the general role of formal processes in mental processes, but I cannot resist calling attention to one massive use-mention confusion implicit in **Fodor's** account. From the fact that, for example, syntactical rules concern formal objects, it does not follow that they are formal rules. Like other rules affecting human behavior they are defined by their content, not their form. It just so happens that in this case their content concerns forms.

In what is perhaps his crucial point, **Fodor** suggests that we should think of the brain or the computer as performing formal operations only on *interpreted* and not just on *formal* symbols. But who does the interpreting? And what is an interpretation? If he is saying that for intentionality there must be intentional content in addition to the formal symbols, then I of course agree. Indeed, two of the main points of my argument are that in our own case we have the "interpretation," that is, we have intrinsic intentionality, and that the computer program could never by itself be sufficient for that. In the case of the computer we make observer-relative ascriptions of intentionality, but that should not be mistaken for the real thing since the computer program by itself has no intrinsic intentionality.

**Sloman & Croucher** claim that the problem in my thought experiment is that the system isn't big enough. To Schank's story understander they would add all sorts of other operations, but they emphasize that these operations are computational and not physical. The obvious objection to their proposal is one they anticipate: I can still repeat my thought experiment with their system no matter how big it is. To this, they reply that I assume "without argument, that it is impossible for another mind to be based on his [my] mental process without his [my] knowing." But that is not what I assume. For all I know, that may be false. Rather, what I assume is that you can't understand Chinese if you don't know the meanings of any of the words in Chinese. More generally, unless a system can attach semantic content to a set of syntactic elements, the introduction of the elements in the system adds nothing by way of intentionality. That goes for me and for all the little subsystems that are being postulated inside me.

**Eccles** points out quite correctly that I never undertake to refute the dualist-interaction position held by him and Popper. Instead, I argue against strong AI on the basis of what might be called a monist interactionist position. My only excuse for not attacking his form of dualism head-on is that this paper really had other aims. I am concerned directly with strong AI and only incidentally with the "mind-brain problem." He is quite right in thinking that my arguments against strong AI are not by themselves inconsistent with his version of dualist interactionism, and I am pleased to see that we share the belief that "it is high time that strong AI was discredited."

I fear I have nothing original to say about **Rachlin's** behaviorist response, and if I discussed it I would make only the usual objections to extreme behaviorism. In my own case I have an extra difficulty with behaviorism and functionalism because I cannot imagine anybody actually believing these views. I know that poeple say they do, but what am I to make of it when Rachlin says that there are no "mental states underlying . . . behavior" and "the pattern of the behavior *is* the mental state"? Are there no pains underlying Rachlin's pain behavior? For my own case I must confess that there unfortunately often are pains underlying my pain behavior, and I therefore conclude that Rachlin's form of behaviorism is not generally true.

**Lycan** tells us that my counterexamples are not counterexamples to a functionalist theory of language understanding, because the man in my counterexample would be using the wrong programs. Fine. Then tell us what the right programs are, and we will program the man with those programs and still produce a counterexample. He also tells us that the right causal connections will determine the appropriate content to attach to the formal symbols. I believe my reply to **Fodor** and other versions of the causal or robot reply is relevant to his argument as well, and so I will not repeat it.

**Hofstadter** cheerfully describes my target article as "one of the wrongest, most infuriating articles I have ever read in my life." I believe that he would have been less (or perhaps more?) infuriated if he had troubled to read the article at all carefully. His general strategy appears to be that whenever I assert p, he says that I assert not p. For example, I reject dualism, so he says I believe in the soul. I think it is a plain fact of nature that mental phenomena are caused by neurophysiological phenomena, so he says I have "deep difficulty" in accepting any such view. The whole tone of my article is one of treating the mind as part of the (physical) world like anything else, so he says I have an "instinctive horror" of any such reductionism. He misrepresents my views at almost every point, and in consequence I find it difficult to take his commentary seriously. If my text is too difficult I suggest Hofstadter read **Eccles** who correctly perceives my rejection of dualism.

Furthermore, **Hofstadter's** commentary contains the

following non sequitur. From the fact that intentionality "springs from" the brain, together with the extra premise that "physical processes are formal, that is, rule governed" he infers that formal processes are constitutive of the mental, that we are "at bottom, formal systems." But that conclusion simply does not follow from the two premises. It does not even follow given his weird interpretation of the second premise: "To put it another way, the extra premise is that there is no intentionality at the level of particles." I can accept all these premises, but they just do not entail the conclusion. They do entail that intentionality is an "*outcome* of formal processess" in the trivial sense that it is an outcome of processes that have a level of description at which they are the instantiation of a computer program, but the same is true of milk and sugar and countless other "outcomes of formal processes."

**Hofstadter** also hypothesizes that perhaps a few trillion water pipes might work to produce consciousness, but he fails to come to grips with the crucial element of my argument, which is that even if this were the case it would have to be because the water-pipe system was duplicating the causal powers of the brain and not simply instantiating a formal program.

I think I agree with **Smythe's** subtle commentary except perhaps on one point. He seems to suppose that to the extent that the program is instantiated by "primitive hardware operations" my objections would not apply. But why? Let the man in my example have the program mapped into his hardware. He still doesn't thereby understand Chinese. Suppose he is so "hard wired" that he automatically comes out with uninterpreted Chinese sentences in response to uninterpreted Chinese stimuli. The case is still the same except that he is no longer acting voluntarily.

**Side issues.** I felt that some of the commentators missed the point or concentrated on peripheral issues, so my remarks about them will be even briefer.

I believe that **Bridgeman** has missed the point of my argument when he claims that though the homunculus in my example might not know what was going on, it could soon learn, and that it would simply need more information, specifically "information with a known relationship to the outside world." I quite agree. To the extent that the homunculus has such information it is more than a mere instantiation of a computer program, and thus it is irrelevant to my dispute with strong AI. According to strong AI, if the homunculus has the right program it must already have the information. But I disagree with Bridgeman's claim that the only properties of the brain are the properties it has at the level of neurons. I think all sides to the present dispute would agree that the brain has all sorts of properties that are not ascribable at the level of individual neurons – for example, causal properties (such as the brain's control of breathing).

Similar misgivings apply to the remarks of **Marshall.** He stoutly denounces the idea that there is anything weak about the great achievements of weak AI, and concludes "Clearly, there must be some radical misunderstanding here." The only misunderstanding was in his supposing that in contrasting weak with strong AI, I was in some way disparaging the former.

**Marshall** finds it strange that anyone should think that a program could be a theory. But the word program is used ambiguously. Sometimes "program" refers to the pile of punch cards, sometimes to a set of statements. It is in the latter sense that the programs are sometimes supposed to be theories. If Marshall objects to that sense, the dispute is still merely verbal and can be resolved by saying not that the program is a theory, but that the program is an embodiment of a theory. And the idea that programs could be theories is not something I invented. Consider the following. "Occasionally after seeing what a program can do, someone will ask for a specification of the theory behind it. Often the correct response is that the program *is* the theory" (Winston 1977, p. 259).

**Ringle** also missed my point. He says I take refuge in mysticism by arguing that "the physical properties of neuronal systems are such that they cannot *in principle* be simulated by a nonprotoplasmic computer." But that is not even remotely close to my claim. I think that anything can be given a formal simulation, and it is an empirical question in each case whether the simulation duplicated the causal features. The question is whether the formal simulation by *itself*, without any further causal elements, is sufficient to reproduce the mental. And the answer to that question is no, because of the arguments I have stated repeatedly, and which Ringle does not answer. It is just a fallacy to suppose that because the brain has a program and because the computer could have the same program, that what the brain does is nothing more than what the computer does. It is for each case an empirical question whether a rival system duplicates the causal powers of the brain, but it is a quite different question whether instantiating a formal program is by itself constitutive of the mental.

I also have the feeling, perhaps based on a misunderstanding, that **Menzel's** discussion is based on a confusion between *how one knows* that some system has mental states and *what it is* to have a mental state. He assumes that I am looking for a criterion for the mental, and he cannot see the point in my saying such vague things about the brain. But I am not in any sense looking for a criterion for the mental. I know what mental states are, at least in part, by myself being a system of mental states. My objection to strong AI is not, as Menzel claims, that it might fail in a single possible instance, but rather that in the instance in which it fails, it possesses no more resources than in any other instance; hence if it fails in that instance it fails in every instance.

I fail to detect any arguments in **Walter's** paper, only a few weak analogies. He laments my failure to make my views on intentionality more explicit. They are so made in the three papers cited by **Natsoulas** (Searle 1979a; 1979b; 1979c).

**Further implications.** I can only express my appreciation for the contributions of **Danto, Libet, Maxwell, Puccetti,** and **Natsoulas.** In various ways, they each add supporting arguments and commentary to the main thesis. Both Natsoulas and Maxwell challenge me to provide some answers to questions about the relevance of the discussion to the traditional ontological and mind-body issues. I try to avoid as much as possible the traditional vocabulary and categories, and my own – very tentative – picture is this. Mental states are as real as any other biological phenomena. They are both caused by and realized in the brain. That is no more mysterious than the fact that such properties as the elasticity and puncture resistance of an inflated car tire are both caused by and realized in its microstructure. Of course, this does not imply that mental states are ascribable to individual neurons, any more than the properties at the level of the tire are ascribable to individual electrons. To pursue the analogy: the brain operates causally both at the level of the neurons and at the level of the mental states, in the same sense that the tire operates causally both at the level of particles and at the level of its overall properties. Mental states are no more epiphenomenal than are the elasticity and puncture resistance of an inflated tire, and interactions can be described both at the higher and lower levels, just as in the analogous case of the tire.

Some, but not all, mental states are conscious, and the intentional-nonintentional distinction cuts across the conscious-unconscious distinction. At every level the phenomena are causal. I suppose this is "interactionism," and I guess it is

also, in some sense, "monism," but I would prefer to avoid this myth-eaten vocabulary altogether.

**Conclusion.** I conclude that the Chinese room has survived the assaults of its critics. The remaining puzzle to me is this: why do so many workers in AI still want to adhere to strong AI? Surely weak AI is challenging, interesting, and difficult enough.

ACKNOWLEDGMENT

I am indebted to Paul Kube for discussion of these issues.

## References

Anderson, J. (1980) Cognitive units. Paper presented at the Society for Philosophy and Psychology, Ann Arbor, Mich. [RCS]

Block, N. J. (1978) Troubles with functionalism. In: *Minnesota studies in the philosophy of science, vol. 9*, ed. C. W. Savage, Minneapolis: University of Minnesota Press. [NB, WGL]

(forthcoming) Psychologism and behaviorism. *Philosophical Review*. [NB, WGL]

Bower, G. H.; Black, J. B., & Turner, T. J. (1979) Scripts in text comprehension and memory. *Cognitive Psychology* 11: 177–220. [RCS]

Carroll, C. W. (1975) *The great chess automaton*. New York: Dover. [RP]

Cummins, R. (1977) Programs in the explanation of behavior. *Philosophy of Science* 44: 269–87. [JCM]

Dennett, D. C. (1969) *Content and consciousness*. London: Routledge & Kegan Paul. [DD,TN]

(1971) Intentional systems. *Journal of Philosophy* 68: 87–106. [TN]

(1972) Reply to Arbib and Gunderson. Paper presented at the Eastern Division meeting of the American Philosophical Association. Boston, Mass. [TN]

(1975) Why the law of effect won't go away. *Journal for the Theory of Social Behavior* 5: 169–87. [NB]

(1978) *Brainstorms*. Montgomery, Vt.: Bradford Books. [DD, AS]

Eccles, J. C. (1978) A critical appraisal of brain-mind theories. In: *Cerebral correlates of conscious experiences*, ed. P. A. Buser and A. Rougeul-Buser, pp. 347–55. Amsterdam: North Holland. [JCE]

(1979) *The human mystery*. Heidelberg: Springer Verlag. [JCE]

Fodor, J. A. (1968) The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy* 65: 627–40. [NB]

(1980) Methodological solopsism considered as a research strategy in cognitive psychology. *The Behavioral and Brain Sciences* 3:1. [NB, WGL, WES]

Freud, S. (1895) Project for a scientific psychology. In: *The standard edition of the complete psychological works of Sigmund Freud*, vol. 1, ed. J. Strachey. London: Hogarth Press, 1966. [JCM]

Frey, P. W. (1977) An introduction to computer chess. In: *Chess skill in man and machine*, ed. P. W. Frey. New York, Heidelberg, Berlin: Springer-Verlag. [RP]

Fryer, D. M. & Marshall, J. C. (1979) The motives of Jacques de Vaucanson. *Technology and Culture* 20: 257–69. [JCM]

Gibson, J. J. (1966) *The senses considered as perceptual systems*. Boston: Houghton Mifflin. [TN]

(1967) New reasons for realism. *Synthese* 17: 162–72. [TN]

(1972) A theory of direct visual perception. In: *The psychology of knowing* ed. S. R. Royce & W. W. Rozeboom. New York: Gordon & Breach. [TN]

Graesser, A. C.; Gordon, S. E.; & Sawyer, J. D. (1979) Recognition memory for typical and atypical actions in scripted activities: tests for a script pointer and tag hypotheses. *Journal of Verbal Learning and Verbal Behavior* 1: 319–32. [RCS]

Gruendel, J. (1980). Scripts and stories: a study of children's event narratives. Ph.D. dissertation, Yale University. [RCS]

Hanson, N. R. (1969) *Perception and discovery*. San Francisco: Freeman, Cooper. [DOW]

Hayes, P. J. (1977) In defence of logic. In: *Proceedings of the 5th international joint conference on artificial intelligence*, ed. R. Reddy. Cambridge, Mass.: M.I.T. Press. [WES]

Hobbes, T. (1651) *Leviathan*. London: Willis. [JCM]

Hofstadter, D. R. (1979) *Gödel, Escher, Bach*. New York: Basic Books. [DOW]

Householder, F. W. (1962) On the uniqueness of semantic mapping. *Word* 18: 173–85. [JCM]

Huxley, T. H. (1874) On the hypothesis that animals are automata and its history. In: *Collected Essays*, vol. 1. London: Macmillan, 1893. [JCM]

Kolers, P. A. & Smythe, W. E. (1979) Images, symbols, and skills. *Canadian Journal of Psychology* 33: 158–84. [WES]

Kosslyn, S. M. & Shwartz, S. P. (1977) A simulation of visual imagery. *Cognitive Science* 1: 265–95. [WES]

Lenneberg, E. H. (1975) A neuropsychological comparison between man, chimpanzee and monkey. *Neuropsychologia* 13: 125. [JCE]

Libet, B. (1973) Electrical stimulation of cortex in human subjects and conscious sensory aspects. In: *Handbook of sensory physiology*, vol. II, ed. A. Iggo, pp. 743–90. New York: Springer-Verlag. [BL]

Libet, B., Wright, E. W., Jr., Feinstein, B., and Pearl, D. K. (1979) Subjective referral of the timing for a conscious sensory experience: a functional role for the somatosensory specific projection system in man. *Brain* 102:191–222. [BL]

Longuet-Higgins, H. C. (1979) The perception of music. *Proceedings of the Royal Society of London* B 205:307–22. [JCM]

Lucas, J. R. (1961) Minds, machines, and Gödel. *Philosophy* 36:112–127. [DRH]

Lycan, W. G. (forthcoming) Form, function, and feel. *Journal of Philosophy*. [NB, WGL]

McCarthy, J. (1979) Ascribing mental qualities to machines. In: *Philosophical perspectives in artificial intelligence*, ed. M. Ringle. Atlantic Highlands, N.J.: Humanities Press. [JM, JRS]

Marr, D. & Poggio, T. (1979) A computational theory of human stereo vision. *Proceedings of the Royal Society of London* B 204:301–28. [JCM]

Marshall, J. C. (1971) Can humans talk? In: *Biological and social factors in psycholinguistics*, ed. J. Morton. London: Logos Press. [JCM]

(1977) Minds, machines and metaphors. *Social Studies of Science* 7:475–88. [JCM]

Maxwell, G. (1976) Scientific results and the mind-brain issue. In: *Consciousness and the brain*, ed. G. G. Globus, G. Maxwell, & I. Savodnik. New York: Plenum Press. [GM]

(1978) Rigid designators and mind-brain identity. In: *Perception and cognition: Issues in the foundations of psychology*, Minnesota Studies in the Philosophy of Science, vol. 9, ed. C. W. Savage. Minneapolis: University of Minnesota Press. [GM]

Mersenne, M. (1636) *Harmonie universelle*. Paris: Le Gras. [JCM]

Moor, J. H. (1978) Three myths of computer science. *British Journal of the Philosophy of Science* 29:213–22. [JCM]

Nagel, T. (1974) What is it like to be a bat? *Philosophical Review* 83:435–50. [GM]

Natsoulas, T. (1974) The subjective, experiential element in perception. *Psychological Bulletin* 81:611–31. [TN]

(1977) On perceptual aboutness. *Behaviorism* 5:75–97. [TN]

(1978a) Haugeland's first hurdle. *Behavioral and Brain Sciences* 1:243. [TN]

(1979b) Residual subjectivity. *American Psychologist* 33:269–83. [TN]

(1980) Dimensions of perceptual awareness. Psychology Department, University of California, Davis. Unpublished manuscript. [TN]

Nelson, K. & Gruendel, J. (1978) From person episode to social script: two dimensions in the development of event knowledge. Paper presented at the biennial meeting of the Society for Research in Child Development, San Francisco. [RCS]

Newell, A. (1973) Production systems: models of control structures. In: *Visual information processing*, ed. W. C. Chase. New York: Academic Press. [WES]

(1979) Physical symbol systems. Lecture at the La Jolla Conference on Cognitive Science. [JRS]

(1980) Harpy, production systems, and human cognition. In: *Perception and production of fluent speech*, ed. R. Cole. Hillsdale, N.J.: Erlbaum Press. [WES]

Newell, A. & Simon, H. A. (1963) GPS, a program that simulates human thought. In: *Computers and thought*, ed. A. Feigenbaum & V. Feldman, pp. 279–93. New York: McGraw Hill. [JRS]

Panofsky, E. (1954) *Galileo as a critic of the arts*. The Hague: Martinus Nijhoff. [JCM]

Popper, K. R. & Eccles, J. C. (1977) *The self and its brain*. Heidelberg: Springer-Verlag. [JCE, GM]

Putnam, H. (1960) Minds and machines. In: *Dimensions of mind*, ed. S. Hook, pp. 138–64. New York: Collier. [MR, RR]

(1975a) The meaning of "meaning." In: *Mind, language and reality*. Cambridge University Press. [NB, WGL]

(1975b) The nature of mental states. In: *Mind, language and reality*. Cambridge: Cambridge University Press. [NB]

(1975c) Philosophy and our mental life. In: *Mind, language and reality*. Cambridge: Cambridge University Press. [MM]

Pylyshyn, Z. W. (1980a) Computation and cognition: issues in the foundations of cognitive science. *Behavioral and Brain Sciences* 3. [JRS, WES]

(1980b) Cognitive representation and the process-architecture distinction. *Behavioral and Brain Sciences*. [ZWP]

Russell, B. (1948) *Human knowledge: its scope and limits*. New York: Simon and Schuster. [GM]

Schank, R. C. & Abelson, R. P. (1977) *Scripts, plans, goals, and understanding*. Hillsdale, N.J.: Lawrence Erlbaum Press. [RCS, JRS]

Searle, J. R. (1979a) Intentionality and the use of language. In: *Meaning and use*, ed. A. Margalit. Dordrecht: Reidel. [TN, JRS]

(1979b) The intentionality of intention and action. *Inquiry* 22:253-80. [TN, JRS]

(1979c) What is an intentional state? *Mind* 88:74–92. [JH, GM, TN, JRS]

Sherrington, C. S. (1950) Introductory. In: *The physical basis of mind*, ed. P. Laslett, Oxford: Basil Blackwell. [JCE]

Slate, J. S. & Atkin, L. R. (1977) CHESS 4.5 – the Northwestern University chess program. In: *Chess skill in man and machine*, ed. P. W. Frey. New York, Heidelberg, Berlin: Springer Verlag.

Sloman, A. (1978) *The computer revolution in philosophy*. Harvester Press and Humanities Press. [AS]

(1979) The primacy of non-communicative language. In: *The analysis of meaning (informatics 5)*, ed. M. McCafferty & K. Gray. London: ASLIB and British Computer Society. [AS]

Smith, E. E.; Adams, N.; & Schorr, D. (1978) Fact retrieval and the paradox of interference. *Cognitive Psychology* 10:438–64. [RCS]

Smythe, W. E. (1979) *The analogical/propositional debate about mental representation: a Goodmanian analysis*. Paper presented at the 5th annual meeting of the Society for Philosophy and Psychology, New York City. [WES]

Sperry, R. W. (1969) A modified concept of consciousness. *Psychological Review* 76:532–36. [TN]

(1970) An objective approach to subjective experience: further explanation of a hypothesis. *Psychological Review* 77:585–90. [TN]

(1976) Mental phenomena as causal determinants in brain function. In: *Consciousness and the brain*, ed. G. G. Globus, G. Maxwell, & I. Savodnik. New York: Plenum Press. [TN]

Stich, S. P. (in preparation) On the ascription of content. In: *Entertaining thoughts*, ed. A. Woodfield. [WGL]

Thorne, J. P. (1968) A computer model for the perception of syntactic structure. *Proceedings of the Royal Society of London* B 171:377–86. [JCM]

Turing, A. M. (1964) Computing machinery and intelligence. In: *Minds and machines*, ed. A. R. Anderson, pp.4–30. Englewood Cliffs, N.J.: Prentice-Hall. [MR]

Weizenbaum, J. (1965) Eliza – a computer program for the study of natural language communication between man and machine. *Communication of the Association for Computing Machinery* 9:36–45. [JRS]

(1976) *Computer power and human reason*. San Francisco: W. H. Freeman. [JRS]

Winograd, T. (1973) A procedural model of language understanding. In: *Computer models of thought and language*, ed. R. Schank & K. Colby. San Francisco: W. H. Freeman. [JRS]

Winston, P. H. (1977) *Artificial intelligence*. Reading, Mass. Addison-Wesley; [JRS]

Woodruff, G. & Premack, D. (1979) Intentional communication in the chimpanzee: the development of deception. *Cognition* 7:333–62. [JCM]