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).