How to turn an information processor into an understander

Aaron Sloman and Monica Croucher
University ol Sussex, Brighton, UK
(Aaron Sloman is now in Birmingham http://www.cs.bham.ac.uk/˜axs)

Commentary on Minds, brains, and programs
John R. Searle
The Behavioral and Brain Sciences (1980) 3, 417-457
http://dx.doi.org/10.1017/S0140525X00005756
Commentary: pages 447-448

Searle's delightfully clear and provocative essay contains a subtle mistake, which is also often made by Al 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 Al 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 scurrying 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 Al 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 physical 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 possibility. 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 operations, 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?