Slightly expanded version of paper in *Proceedings AISB Conference, Warwick University 1985* published as:
A.G. Cohn and J.R. Thomas (eds) *Artificial Intelligence and Its Applications,* John Wiley and Sons 1986.

DID SEARLE ATTACK STRONG STRONG OR WEAK STRONG AI?

Aaron Sloman
School of Cognitive and Computin Sciences
University of Sussex
Brighton BN1 9QH

ABSTRACT

John Searle's attack on the Strong AI thesis, and the published replies, are all based on a failure to distinguish two interpretations of that thesis, a strong one, which claims that the mere occurrence of certain process patterns will suffice for the occurrence of mental states, and a weak one which requires that the processes be produced in the right sort of way. Searle attacks strong strong AI, while most of his opponents defend weak strong AI. This paper explores some of Searle's concepts and shows that there are interestingly different versions of the 'Strong AI' thesis, connected with different kinds of reliability of mechanisms and programs.

Keywords

Searle, strong AI, minds and machines, intentionality, meaning, reference, computation.

Introduction

John Searle's Reith Lectures broadcast on BBC radio in 1984, and subsequently issued as a book by BBC publications, have attracted a lot of attention, in particular his attack on the Strong AI thesis, which claims that suitably programmed computers would understand, think, feel, etc. I do not believe that published commentaries have dealt adequately with Searle's arguments. Moreover, I believe that although his main conclusions are quite wrong, he raises issues which are potentially of considerable importance for AI, and that his arguments should therefore not simply be ignored by those who disagree.

A more detailed version of the argument was originally published in 1980, and subsequently reprinted in [Haugeland 1981] and [Hofstadter 19??]. The 1980 journal article was published alongside criticisms by a motley collection of commentators, with Searle's reply. Since it provides the most complete discussion, I shall refer to the 1980 collection in this essay. I shall try to show that the strong AI thesis, as defined by Searle, has two interpretations, one of which is stronger than the other. In fact there is a range of versions of varying strengths. I do not believe that many of those who thought they were disagreeing with Searle would in fact wish to defend the strongest strong AI thesis. However, close examination suggests that that is the one Searle thought he was attacking, though he, like almost everyone else, failed to be clear about this.

There is an extreme version of the Strong AI thesis against which his attack is successful. But this leaves open the status of a weaker version, which is probably the one his commentators thought they were defending (insofar as they were clear about what they were defending). This paper is mainly concerned with clarifying the problems. A weak version of the Strong AI thesis is defended at length in [Sloman 1985].

Although Searle's discussion is relevant to mental concepts in general, he focuses on the concept of 'understanding', partly because it is often claimed that because computers can understand symbols (e.g. in the form of instructions), we thereby have a basis for designing them so that other types of mental states can occur, such as reasoning, perceiving, planning, deciding, intending, believing, thinking, imagining, etc. He rightly selects 'understanding' as the most plausible mental concept to attribute to machines. If he can demolish the most plausible example, the rest collapses.

Occasionally he uses phraseology which suggests that computers cannot manipulate symbols which have meaning. This is not exactly what he means. He agrees that they have meaning *for us.* That is, he claims that their meaning is 'derivative' or 'secondary' for it is we, not the machines, that understand the symbols. Exactly what it means to say that someone or something understands symbols is not at all clear, and I have discussed this at length in [Sloman 1985], sketching some global design requirements for machines to understand. The present paper is concerned with a different issue, namely the types of causal powers required in a machine which can have mental processes.

Minsky, in his commentary on Searle [Searle 1980], alleged that concepts of ordinary language are not sufficiently powerful to be used for posing interesting questions in this area, or formulating insightful answers. Like Minsky, I don't think our ordinary concepts, like 'understand', 'conscious', 'intend', 'cause', are subtle and precise enough to cope with all interestingly different cases. But there are clear differences between the main sub-cases, whether or not we can label them accurately in English. Approximate labels will do for now. Eventually they will need to be replaced by more precise ones, defined over a wider range of cases. For now, colloquial language is a rich and powerful source of cues (see [Sloman 1978] - chapter 4), and it is essential that we use it as the best starting point we have for these discussions. Later work should define more useful technical concepts.

Different Strong AI theses

Searle introduces the 'Strong AI' thesis as stating that an appropriately programmed computer 'really is a mind', 'literally has cognitive states' (page 417, Searle 1980). He elaborates by saying that having cognitive states, or mental processes, involves intentionality: i.e. there are states which involve some semantic content, that is some reference to objects, properties, or relationships outside themselves. The thesis does not claim that machines can already have mental states, only that they would if suitably programmed. What sorts of programs would be suitable is still an unsolved research problem.

The strong AI thesis implies that the existence of certain still unspecified computational processes would be *sufficient* for the existence of intentionality, sufficient for the existence of states and processes with a meaningful content. So, according to the Strong AI thesis, there is some description C of a type of computational process such that if C is true of any machine, then that machine has mental states (e.g. can understand symbols, or use them to refer to other things). This can be summarised formally:

T1. C(x) -> M(x)

Searle is happier with the 'weak' or 'cautious' AI thesis which merely claims that the computer is a useful tool in the study of mind. He contrasts the strong thesis that mind could be replicated in a computer with the weaker thesis that it could be simulated. He seems to believe that mental processes can be simulated in computers, as closely as required, but that they will never thereby by replicated, just as a meteorologist's simulation of a rain-storm, however detailed, does not actually produce wetness in the machine. In both cases the simulation may play a useful role in developing and testing theories.

However, unlike some who accept this 'Weak AI' thesis, Searle apparently does not believe that the concepts of AI or computer science will play any useful role in psychological theorising, any more than they do in theorising about the weather. (I owe this comment to Margaret Boden.) I shall not discuss the 'weak AI' thesis any further.

The 'Strong Calculator Thesis'

It is interesting to compare the Strong AI thesis with a 'Strong Calculator Thesis', which may be found plausible by some who reject the strong AI thesis. Suppose you produce a machine whose behaviour maps onto algorithms for doing arithmetical calculations. Does it calculate, or does it merely *simulate* calculation? The Strong Calculator thesis states, quite plausibly, except for a qualification noted below, that if the behaviour of a calculator is simulated in sufficient detail, including the internal operations, then its behaviour is actual calculation, i.e. you have another calculator. That is, calculation cannot be simulated without being replicated. Are mental processes like this?

Different answers are possible, for instance, that ALL are, that SOME are, and that NONE are. Searle thinks none are.

Does instantiating the right program produce understanding?

I shall now show that there are different interpretations of the 'Strong AI' thesis, each taking a form something like T1 above, where C(x) asserts that some sort of computational process occurs in x, and M(x) that some kind of (intentional) mental process occurs in x. I shall not attempt a general definition of 'computational', any more than Searle does, though like him I shall assume that in a computational process there is some form of behaviour, B, which 'instantiates' a formal specification, or program, P. Searle uses this notion of 'instantiation' without analysis, and since a number of people have found it puzzling, I shall indicate what I think he must have meant.

In general, a program P specifies a set of possible transformations of a ’virtual machine’ which is defined to have various components which may change their states in certain limited ways, e.g.: variables may have values which change, sub-processes may be invoked and return results to the calling process, structures may be created and given certain contents, etc. To say that a physical process instantiates a program, then, is to say that there is some way of characterising the process such that it maps onto one of the permitted classes of processes in the relevant virtual machine. In general, this will involve abstracting from many details of the physical process, and it may even be necessary to go through various transformations of virtual machines to demonstrate that a program in a very high level language is instantiated in a particular physical machine. Such complications will be ignored in what follows. Instead we shall merely assume that saying that x has a computational process is equivalent to saying that the behaviour of x ’instantiates’ some program P, in the sense loosely indicated above. (Some of the behaviour will be internal.)

In the light of all this, we can expand the Strong AI thesis, T1 above, into the assertion that there is some program P, as yet unspecified, such that x’s behaviour instantiating P is a sufficient condition for x to have mental states.

T2. Instantiates(P, B(x)) -> M(x)

Alternative formulations of the left half will later be used to distinguish weaker and stronger versions of the Strong AI thesis.

In a series of remarks embedded in his 1980 papers Searle reveals that he interprets the strong thesis in such a way that no very strong connection is required between P and the behaviour, as long as the behaviour produces a formal pattern of structure manipulations of an appropriate sort. E.g. he talks about ’operators on purely formally defined elements’ (p.418), ’the formal shadow’ cast by brain processes not being enough for intentionality (p.422). He refers to Weizenbaum as showing how to construct a computer (sic!) using a roll of toilet paper and a pile of stones (p.423), not noticing that something else is needed to move the stones about.

I don’t claim that Searle understood enough to have any very precise notion of computation in mind, and it is likely that he slid unwittingly between the different versions I shall distinguish. But for the purposes of his argument, and his more rhetorical flourishes, he seemed to use a very limited characterisation. I.e. he takes ’strong AI’ to state that the mere existence of some formally defined pattern of processing, or the mere instantiation of a program (’the right program of course’ p.422) would be sufficient for the occurrence of some mental process, such as understanding a Chinese text.

To refute this strong AI thesis he claims that the mere instantiation of a program cannot suffice for the production of mental processes, and he backs this up this by arguing that whatever the computational processes might be that are supposed to suffice for understanding of a Chinese text, he can replicate them in a situation where there is no understanding of Chinese, though there may be external behaviour indicative of such understanding. Only a behaviourist, he claims would regard the external behaviour as proof of understanding.

He then suggests that the reason why mere instantiation of a program cannot produce mentality is that mental processes involve intentionality, and only an underlying machine with the right causal powers, e.g. the brain, can produce intentionality, though he doesn't say how it can do this. Some of his wording suggests that he thinks of intentionality as a kind of substance, like milk, produced by a chemical process, like lactation. But what he is really saying (e.g. p.451 'I do not, of course, think that intentionality is a fluid') is that the existence of certain brain states or processes is *sufficient* for the existence of certain intentional (mental) states or processes. This claim would be acceptable to most defenders of strong AI. But Searle also claims that there are certain neurological conditions N such that

N1. N(x) is necessary for M(x)

where N refers to a collection of causal powers to be found in brains, but which do not come into existence merely because any purely formal pattern of processing occurs.

He does not actually produce any argument for N1, and he is totally vague about the precise nature of N, apart from allowing that in principle N might be satisfied by some electronic artefact, provided that it was sufficiently similar in its powers to the brain. Neither does he give any reasons why N should be necessary for M. However, these gaps do not affect the validity of refuting T1 by producing an example where C(x) is true but not M(x). The same example refutes the more precise T2.

Searle's refutation of T1 and T2

His example is well known: assume there's some program Pc supposedly adequate to enable a computer to understand Chinese. Then T2 asserts that if x's behaviour instantiates Pc, then x understands Chinese, and therefore mental states occurin x.

Tc. Instantiates(Pc, B(x)) -> Uc(x)

Searle claims that provided Pc is expressed in a programming language he understands, he can take the place of the computer and produce an instantiation of Pc which will fool others outside the room into thinking that something inside the room understands Chinese (if Pc is a good program). But, since he doesn't understand Chinese, and (he claims) the total system consisting of him executing the program doesn't understand Chinese, it follows that the alleged mental state does not exist. So C(x) (that is B(x) instantiating Pc) is not sufficient for M(x) in this case.

His logic is unassailable, if his premisses are true, in particular the premiss that nothing understands Chinese in this situation.

His original paper, and the reply to commentators, both attempt to dispose of obvious objections to this argument, though parts of the discussion consist of mere assertion and counter-assertion. He refutes various fairly obvious replies by elaborating the example in various ways, including connecting TV cameras and external limbs to the room, and completely memorising the program so that everything goes on inside his own head. He claims that there is no understanding of Chinese produced by the running of Pc, because no matter how the example is varied he experiences no understanding of Chinese. He seems to rely on introspection as sufficient to establish this, suggesting that the only alternative to introspection is to use behaviourist criteria for understanding, and, like many philosophers and psychologists he regards behaviourism as already refuted elsewhere.

Four years later, his Reith lectures show that his opponents have not made him shift his ground at all. I suggest that this is because there is a very strong interpretation of T1, T2 and Tc, in which they are false, and which, he has refuted. There is no published explicit defense of this very strong AI thesis, as far as I know. I shall later explain why it is implausible. The reason counter-arguments have not affected him could be in part that at best they defend a weaker version of the strong AI thesis, i.e. one which proposes more specific conditions as sufficient for mentality.

Searle's interpretation of T1, T2 and Tc is a very strong one. He makes it clear that on the view he is attacking, it makes no difference exactly how the behaviour is produced, as long as it instantiates the right program: as long as it has the right *formal* properties, it is supposed to suffice for the existence of mentality. However, by contrasting different relationships between program and behaviour I shall show that this very strong AI thesis is not plausible, and that refuting it leaves more plausible versions unscathed.

Causal links between program and behaviour

So far all we have required for C(x) to be true is that x's behaviour instantiates the relevant program P. In the general sense of 'instantiate' defined above, this is a very weak relationship between P and B. For instance, there need be no causal connection between any representation of the program and the behaviour. The instantiation might be purely fortuitous, like the wind blowing leaves about to form what looks like an English sentence, or a numerical calculation.

Consider again the 'Strong Calculator Thesis'. If a physical process just happened to 'instantiate' some algorithms for numerical calculation, but did so purely by chance, like the leaves in the wind, would this be calculation? Does the strong calculator thesis apply only to behaviour produced by some appropriate causal mechanism? If so, this may give us a clue as to what, if anything, is correct in Searle's argument.

It may seem at first as if there is a purely semantic issue here: a mere verbal question about the scope of a word, which we may in fact define as broadly or as narrowly as we like. However, more is at stake. If we think of a calculator as having a function, then it is not just a verbal question whether something has a mechanism which can be trusted to perform that function *reliably.* Irrespective of what we decided to *call* a calculation-like pattern of motion produced by the wind, we would be foolish to *rely* on it for designing bridges or aircraft. It is for good reasons that computer manufacturers go to a lot of trouble to make the machines which run the programs reliable. In particular, they are designed so that there is a strong causal connection between physical representations of programs and the behaviour they produce. Mere instantiation is not enough.

Similarly, in the design of an intelligent system, it is important that mechanisms be used which can be *relied* on to produce coherent behaviour. If mere instantiation of the right programs were enough for the production of mental states, then we could argue that all mental states are to be found in thunderclouds. The argument is analogous to the argument which states that there are letters on the surface of a blank sheet of paper, though nobody has marked their outlines. The portions of the surface exist whether physically marked or not. Similarly, all Shakespeare's sonnets existed on the walls of ancient Greek houses, though nobody knew they were there, because the letters had not been marked out. Just as portions of a homogeneous surface can be thought of as instantiating letter shapes, so can portions of a very large mass of randomly moving particles be thought of as instantiating a program, though there is no physical demarcation of the various sub-process and sub-substructures.

So, if the strong AI thesis is defined as claiming that ANY instantiation of the right program will suffice for the production of understanding, then that thesis implies that the process of understanding Chinese exists undetected in a thunder cloud, along with many other mental processes.

This shows that the extreme strong AI thesis is absurd, and I do not believe that any of Searle’s opponents would have bothered to reply if they had appreciated that he was attacking this thesis, though I have met one or two intelligent people who have been tempted, though only for a short time, to try to defend it, along with the consequence that thunder-clouds have thoughts.

It is important to remember that this is not a purely verbal issue. A ’cloud-mind’ embodied in this way would lack the causal basis required for interaction with other minds, for perception of an external world, for control of actions. Moreover, the different portions of the mind, like a store of beliefs, a store of motives and a planning mechanism would not interact in any reliable fashion. The beliefs and motives would not *produce* goals or plans, any more than a message embedded in the surface of a wall without any physical demarcation can communicate.

A less extreme version would require physical demarcation of whatever instantiates the program. However even this is not enough to guarantee that the process can be depended on, if it is allowed that the behaviour might be produced by chance, rather than by a causal mechanism related to the program.

So not everything which formally instantiates a program can fulfil the functions which the running of that program may be supposed to fulfil.

Can a person function as a computer?

We can now return to Searle’s Chinese-understanding experiment. He has assumed that for the Strong AI thesis it does not matter how the behaviour is produced, so long as it instantiates an appropriate program. So it doesn’t matter whether the instructions are followed by a computer or a person or even whether the corresponding processes are produced by some purely random process. But we now see that it is important to distinguish different cases. In particular, the behaviour in a computer does not merely instantiate a program; rather a physical representation of the program *causes* the behaviour to occur, with the aid of circuitry (and software in some cases) carefully designed to ensure that there is a reliable connection between program and behaviour. Any engineer designing an intelligent system would require something like this. Evolution likewise.

Understanding, and other mental states and processes, are not isolated happenings. They are capable of being produced by certain sorts of things, and they are capable of producing certain sorts of effects. They form part of a rich web of causal connections. For instance, understanding what someone is saying is capable of interacting with your desires, hopes, fears, or principles, to produce new desires and actions. An intelligent system requires ’motive generators’ capable of reacting to new information by setting up new goals. (Sloman 1978, Sloman and Croucher 1981). A person could not be said to understand English, if there were no way his ’understanding’ of a sentence like ’Your house is on fire’ could make him want to do anything. Structural conditions for understanding might be satisfied, but not the functional conditions, just as an accelerator pedal detached from a car cannot produce any acceleration.

So, instead of T2 and Tc, we need to modify the strong AI thesis to assert, in its general form, that behaviour reliably caused by appropriate programs and capable of causing (or influencing) certain other mental states will be a mental process. Very schematically:

T3. (P => B(x) & Possible(B(x) => M’(x)) -> M(x)

Where ’=>’ represents some form of causation, and M’(x) refers to some other mental state (or collection of states), distinct from M(x). ’->’ is logical sufficiency rather than causation. This may appear to be circular because of the occurrence of a ’mental’ predicate before and after ’->’. However, it need not be a vicious form of circularity. T3 can be interpreted as part of a specification for a *network* of interacting processes including various kinds of mutual causation, like feedback loops.

But doesn’t this modified version of the Strong AI thesis still collapse if Searle is right in saying no understanding is produced when he carries out the instructions in the program. For his behaviour is caused by the program, and the program might consist of several interacting sub-programs one of them being concerned with understanding and others with the formation of beliefs and desires. Thus the left half of T3 would appear to be satisfied, yet Searle claims there is no understanding of Chinese, no matter what is in the program P.

One form of reply would assert that there would be a process of understanding, despite Searle’s inability to introspect it. (He takes it as obvious that he would not experience understanding. Actually, since the whole experiment is hypothetical we don’t even know for sure that he wouldn’t, as a result of achieving fluency with the program, come to feel he understood Chinese as well as English. But let that pass.) Since many do not find behaviourist criteria adequate, something more than behavioural evidence would have to be invoked. I have attempted this in [Sloman 1985] which sketches some of the design requirements for a machine to understand symbols in the way we do, and then attempts to show that a substantial subset is already satisfied by simple computers. The paper also argues against attempting to use ordinary concepts, like ’understanding’, to draw global distinctions in the space of possible behaving systems.

However, even if we can demonstrate that something very like human understanding would occur in a machine driven by an appropriate program, it would not follow that the same process would occur if Searle took the place of the computer. There really is a difference between a computer obeying instructions in a program, and a person doing so. If, as Searle thinks, there would be no understanding in his hypothetical system, this may be (partly) accounted for by the fact that the kind of causal link between program and behaviour in a computer is very different from the kind of link between written or memorised instructions and a human being acting like a computer. Once again, the notion of ’reliability’ is important. Roughly speaking, in a computer the program *drives* or *controls* the process, whereas in the case of a human being it merely *guides* the process.

As ever, the labels we use are not the important thing. What is important is the degree of reliability of the connection between program and behaviour. Informed people would not wish to fly in a plane whose automatic landing system used a person in a box, rapidly obeying instructions in place of the computer for which the instructions were originally written. (This is on the assumption that the person gets no more data than the computer now does.) The point is not that the person would not be as fast as a computer, but that in the case of human beings there is a relatively loose and unreliable linkage between instructions and behaviour. All sorts of different things can, and do, interfere, for instance, boredom, sleep, distractions, or a dislike of the instructor. The human mind is designed (whether by evolutionary processes or God doesn't matter for our purposes) to be enormously flexible and responsive in the pursuit of its interests and goals, and that very characteristic makes it not very suitable as a processing element in a larger system which has its own interests and goals. The same would apply to a really intelligent robot. This is part of the moral of Huxley's Brave New World and stories in which robots revolt.

A hypnotised person would be an interesting intermediate case. But we understand so little about hypnosis and how it differs from many other cases of persuasion and coercion that it is probably not worth discussing in this context. Another intermediate case would be a person forced at gunpoint. Even the latter can be argued, following existentialist philosophers, to be a case where the combination of gun and program do not force the person to obey: for heroism, obstinacy, impulse, and many other possibilities can intervene.

What's Right in Searle's Argument

I conclude that Searle is right to suggest that, as far as our ordinary concepts of calculating and understanding are concerned, not just any production of patterns fitting an (appropriate) program specification is enough to produce understanding. He is right to suggest that the processes must be produced by a system with the right causal powers. This is a requirement not only for mental processes, but for any computational process which is used as a reliable basis for taking decisions or controlling anything. The mere fact that a pattern of leaves or water molecules happens to solve a mathematical problem, does not make either the leaves or the water into a calculator.

However, if this is the upshot of Searle's argument, I doubt whether he or any of his opponents would have bothered arguing about it, had they seen this clearly.

Varieties of Program/behaviour relationships.

Whether or not the debates were at cross-purposes, an interesting notion has emerged, namely that there may be different relationships between program and behaviour. It is worth exploring some of the differences, as they are relevant to real problems concerning the design of intelligent systems. We can analyse the relationships in terms of the engineering requirement for reliability.

Within conventional computing we find several different sorts of relationships between a program and the behaviour which instantiates it. For instance, very often there is no direct causal link between a representation of the program P and the behaviour, for the simple reason that a compiler is used to translate P into a quite different program in the computer's machine language, and only the new program is stored in the computer's memory and causally linked to the behaviour. There is a causal connection between the original program and the behaviour, but it is fairly remote. Nevertheless, considerable effort is normally put into ensuring that both the compilation process and the subsequent processes involve tight causal links with no scope for unpredictable deviations. However, compilers can have bugs. Moreover, there is often some human intervention in that compiling and running may be quite different processes, both initiated by a person. Normally the mere existence of the original program in a computer's files does not produce any behaviour, without such human intervention.

A somewhat different case arises when a physical representation of the program P is stored in the memory of the computer and directly interpreted either by hardware or software. Here the causal link between program and behaviour is much more direct, and less open to human error. Hardware faults are still possible, but in (good) modern computers the frequency with which things go wrong is very small indeed.

However, anyone familiar with computing systems will by now have noticed that my claim that computers are designed to produce tight connections between program and behaviour needs qualifying, even for machine-code programs in the computer's memory. Computer operating systems need to be able to control individual programs, and therefore the latter cannot be given complete autonomy, and their ability to 'drive' the computer therefore needs to be limited. For instance, the program will not be allowed to access portions of memory or disk files for which it lacks the 'privilege'; and, if the machine is shared between different users, individual programs will not be able to run indefinitely. The operating system, with the aid of suitable hardware, can normally interrupt and either suspend or abort a process if it takes up too much time and other processes are waiting. Often even a human user can interrupt the process by pressing appropriate keys on the keyboard.

So the operating system's ability to interrupt a running program for all sorts of different reasons is similar to the ways in which a human being may be unreliable at executing a program. Does this mean that an AI program running under such an operating system would not really produce mental processes?

There are many different sorts of cases to be distinguished, and trying to use concepts of ordinary language to draw a definite boundary between the different cases is pointless. Rather, we need to analyse what the differences are and what their implications are. In particular, we need to distinguish potential gaps in the causal chains which simply reduce reliability from those which increase reliability relative to some more global function.

Levels of reliability in intelligent systems

The operating system's ability to interfere with running programs reduces, in a sense, the reliability of individual programs, but increases the reliability of the whole machine in relation to such goals as protecting privacy, sharing resources according to some fair scheme, and preventing faulty programs corrupting other processes or files.

An intelligent robot, comparable in complexity to a human being, would require a similar relationship between low-level actions and higher-level goals. The main reasons for this are (a) the environment will always be partly unpredictable, (b) new goals can emerge at any time, (c) dealing with new developments may involve time constraints. (a) implies that no matter how carefully plans are made, it will often be necessary to modify them in the course of execution because of some unexpected new discovery. (b) implies that new motives may have higher priority than, and be inconsistent with, the motives responsible for current actions, which may therefore need to be aborted or suspended. (c) implies that it will often be necessary to interrupt actions, or even thought processes, in order to deal with new information. (The arguments are spellt out in more detail in [Sloman 1978, chapter 6], [Sloman and Croucher 1981], [Croucher 1985]).

These considerations lead to a requirement for intelligent systems to have a complex computational architecture which allows different processes to run asynchronously, for instance, plan-execution, perceptual monitors, reasoning and decision-making. Assuming that all the different sub-processes are produced by executing programs of some sort (not necessarily programs of types which we now know how to write), we clearly require that the low level machinery be capable of running the programs reliably. However, we also need chinks in armour of the causal chains to allow low level processes to be modified in the light of higher level needs and new information. So, in terms of the low level programs, the machinery must not be totally reliable. Knowledge of a low level program cannot provide a totally secure basis for predicting its behaviour, if something else can interrupt or modify it.

This sort of organisation will then permit a more global reliability to be achieved, so long as the facilities for one process to modify or interrupt another are part of a coherent design, in which everything is ultimately driven by the goals of the system. But for this, a system will not be intelligent. For instance, it would tend to thrash about pointlessly.

So we now see another difference between what is required for an intelligent system and the sort of situation envisaged by Searle. In his system there will be not only the goals and beliefs of the system represented by the AI program which he is executing, but also his own goals and beliefs; and there is no reason to assume that they will form the sort of coherent whole required for the design of a mind. Thus, even if a process partly analogous to understanding occurs, it will not be capable of fulfilling the functional role of understanding, namely reliably providing input to the belief system, the reasoning and decision-making processes. It may do this part of the time, but only so long as Searle's goals and decisions do not intervene. So we do not have the right sorts of causal links in the total system.

Admittedly, this is not as extreme as the case of processes instantiated purely notionally in a thunder cloud, or processes produced by the wind blowing leaves about without being driven by any program. Nevertheless it is different from the case of an integrated intelligent system, and the difference could be important to anyone who attempted to use Searle's 'machine' as a servant or even as a friend or confidant.

So the causal powers of the different sorts of processes instantiating the same program might be very different, depending on how the processes were produced. Mental processes are distinguished and constituted by their causal powers, including their ability to influence other mental processes. (This idea is developed in Ryle [1949].) A process is not a planning process if it does not dependably issue plans which bear a systematic relation to goals, constraints, available beliefs, etc. A process is not a reasoning process if it cannot be relied on to produce conclusions which are in some sense implied by the premisses. A perceptual process must produce reliable percepts most of the time. (Human visual illusions are themselves systematic and predictable.)

All these considerations suggest that only a weak version of the strong AI thesis should be taken seriously. A weak version which requires programs to be *in control* of processing, is not refuted by any thought experiment involving a process in which the programs are not in control but perhaps, at best, guide the behaviour of an intelligent person, like John Searle.

Why Searle thought anyone would wish to defend the stronger versions of the strong AI thesis is not clear, though perhaps he wasn’t sufficiently precise about his own arguments to have attributed this extreme view to anyone.

Various commentaries on the original, including mine, allowed (mistakenly, I now think) that even when the computer is replaced by a person, the resulting processes would be mental ones, claiming to disagree with Searle’s intuitions about this case. Searle responded that it had nothing to do with intuitions, since the person following rules clearly did not understand Chinese, and neither did the total system of person, books of rules, pencil paper, etc. As already remarked, in order to avoid behaviurism, he felt constrained to use introspective criteria for saying this. Others use behavioural criteria. The debate has an air of pointlessness.

Recapitulation

We have seen that the condition C(x) mentioned in the strong AI Thesis, T1 may be more or less stringent. If more stringent (i.e. harder to attain) then the thesis is weaker. If C(x) is easy to attain, then we have a very strong thesis saying that ’mere’ C(x) is sufficient for M(x). An extreme version is T2 which merely requires that the behaviour of x instantiate some program P. A weaker thesis asserts that a tight causal connection between P and the behaviour is sufficient for M(x).

We have begun to explore a range of cases in which more or less strong connections between program and behaviour are required. In particular, causal links are necessary if the system is to be able to fulfil any functions. At a low level relatively strong causal connections are required between program and behaviour, but they should not be totally reliable, so that detailed processes can be controlled in accordance with higher level needs and new information. When a program P is running on a computer where another program, O, the operating system, has the power to interrupt, suspend, or abort the process, P is not fully in control but the combination of P and O is.

An intermediate case is Searle’s example: an intelligent system with its own goals, reading the instructions in the program, and obeying them. We can label this as the program ’guiding’ the behaviour. A system built out of such units will lack the coherence required for intelligence, since the potential extraneous influences in the form of the processor’s own goals and beliefs will undermine the reliability of the subsystem in performing the functions implied by words like "perceive", "understand", "deliberate", etc.

Some of the commentators in [Searle 1980] (e.g. Block and Fodor) did notice that there might be significant differences between different ways in which behaviour instantiating a program might be produced, though none pointed out the consequence that there are different interpretations of the strong AI thesis, depending on the type of control the program is required to exercise.

We thus have different interpretations of strong AI depending on which sort of relation between program and process is alleged to suffice for M. I believe Searle’s opponents (often unthinkingly, as they had not noticed the distinctions) were defending a weak version of Strong AI, in which the program must be in control in order to produce M. I believe Searle was actually attacking the strongest version, in which any instantiation, no matter how produced is alleged to suffice for M. And he managed to confuse himself and almost everyone else by producing an example of the middle kind where the program merely *guides* the behaviour of an intelligent agent. This was confusing because the phraseology of his text shows clearly that he thinks of strong AI as claiming that ’mere’ formal patterns, or any instantiation of an appropriate program, would suffice for M. He uses the middle case because, in order to avoid the ’other minds’ problem he relies ultimately on the introspections of the human computer (i.e. himself) as a guarantee that there is no understanding of Chinese. He appears to regard it as essential to conduct the argument in ’first person’ terms because only behaviorists could regard external behavioural criteria as adequate to settle the question whether (e.g.) understanding has occurred, and he takes it for granted that behaviourism is a dead horse.

Why different causal powers matter

If a suitably programmed computer takes in some numbers and then prints out their sum then we have a process which fits the specification of the program driving the computer. But suppose the wind in some forest were to blow leaves about so that first we had leaves forming the shapes of two numbers, and then, after certain additional blowing about, a pattern of leaves corresponding to the sum of those numbers. Would we want to call that calculating? The ’strong strong calculator thesis’ would say that ANY instantiation of the program would suffice for the production of a calculation. A ’weak strong calculator thesis’ would say that only if the processes are controlled by the program will that suffice for the processes to form a calculation.

A good principle in such debates is not to argue about where to attach labels, but rather to understand the differences between different options. We do not need to squabble over conflicting intuitions as to how to use the word ’calculation’, for, no matter how we label them, there are objective differences between the two cases. In particular, one of them would be a reliable basis for taking engineering decisions and the other would not. (If the wind and leaves invariably produced the right result when someone shouted out an arithmetical problem, then this would strongly suggest that the process was not random, and that some causal mechanism existed which could account for the reliability.) Similarly there would be a difference between using a properly programmed electronic calculator and using a person guided by the program. The person would be subject to a much richer variety of possible interrupting, diverting, or distracting influences, including purely internal influences like suddenly remembering an unfinished task, or hating the person who asked the arithmetical question.

These objective differences affect the possible role the different sorts of calculation could usefully play in our practical activities. Likewise, apparently intelligent behaviour, instantiating certain programs, might play different sorts of roles in social interactions or even in the purely mental interactions in the mind of a robot. For instance, a robot whose planning processes were subject to the quirks of an intelligent sub-agent interpreting its planning programs, might often form plans which undermined its own goals. A social robot, whose programs were not in control of its behaviour, would not be a reliable servant or friend. In the limiting case, a robot whose purported processes of perception, inference, planning, etc. were not generally under the control of its programs, could not have anything like a unitary mind. It might be more like a deranged person than a rational agent. Nobody would wish to entrust it with any important task. If for a time its processing just happened to comply with its programs it would not suddenly become coherent. It would merely appear to be so, but would still not be able to play a dependable role in social relationships e.g. as servant or friend.

None of this proves that the weak strong AI thesis is correct. All I have done is show that it survives Searle's attack. I have come close to saying what Searle himself said about the brain (though in a quite different spirit), namely that not just any old set of processes satisfying some formal condition will suffice for the existence of intentionality. The processes must be produced by some mechanism 'with the right causal powers'. Unlike Searle, however, I have not said that having the right causal powers depends on having the powers of a human brain. Brains have an adequate set of causal powers, but there might be other adequate mechanisms which work in a quite different way. It is an empirical question whether any physical system other than a naturally produced human brain is capable of supporting the full set of processes required for mental states of a human type.

Two kinds of causal powers

I have agreed with Searle that for a computational state or process to be an example of understanding, or some other mental state, the processes must be produced by a mechanism with the right causal powers. He should also have stated that the processes must themselves have the right causal powers, since mental processes are capable of interacting with arbitrarily many other mental processes. I have argued (following Ryle [1949]) that part of what defines various sorts of mental states and processes is their causal relations with other mental states and processes. (This is the truth that behaviourists nearly discovered!) So, part of what makes a process perception is its capacity to change beliefs and play a role in the guidance of actions. Part of what makes a process reasoning is its ability to influence motives as well as beliefs. Part of what makes a process understanding of Chinese is its ability to generate new beliefs and motives. And so on.

A full discussion would require an analysis of the global architecture of a mind. There are different possible architectures for minds of different sorts - e.g. it is certain that some animal minds are simpler than ours. Searle seems to think that AI is committed to a monolithic model of a mind as constituted by a single machine running a single giant program. We have barely begun to understand the range of possible computational architectures, and it is naive to suppose that current AI research is relevant only to systems with existing global organisations, even if some AI workers are that naive.

Thus there are two sorts of causal powers constitutive of mental processes. First there are the mechanisms which actually enable programs to 'drive' or 'control' processes. Secondly the symbol manipulations themselves must be embedded in a network of causal relationships, so that they are capable of influencing or being influenced by other mental processes. In a complete discussion it would be necessary to elaborate on this by showing how processes using any one mental capacity are intrinsically capable of affecting other processes. For instance, perceiving something could in principle produce new desires, thereby producing a planning process, and later an action. This web of dispositions to interact is very complex in human beings, and by exploring different organisations we can begin to build up a theoretical understanding of the space of possible behaving systems.

Searle would, of course, reply that no matter how complex the network of interacting processes reuired to produce a Chinese understander might be, it could be simulated on a massive computer made of lots of ignorant English-speakers, without creating any understanding of Chinese. If he is right (a point on which some critics don't share his intuitions - for that is all they really are), then that is at best a refutation of a relatively strong strong AI thesis, since, as we have indicated, that sort of program instantiation would differ in important respects from an integrated design, involving no potentially subversive intelligent sub-components.

Primitive forms of meaning

Searle's attack on the strong AI thesis seems to have been based in part on his limited understanding of the nature of programs and computers. He appears to believe that in some important way programs are purely syntactic objects, that there is nothing remotely like intentionality in the sorts of processing that most computers do, except for 'derivative' intentionality which documents in a filing cabinet also posses because WE interpret them as meaningful. He argues, especially clearly in the Reith lectures, that since semantics cannot be developed out of pure syntax, it is impossible for a mere programmed computer to have mental processes.

It is very interesting that one of Searle's main premises is just wrong, even though it is widely believed. In particular, it is not true that programs are purely syntactic objects, and that all their semantics is derivative on our reading things into them.

This is not true because there are important ways in which even some of the most primitive programming languages have quite a rich semantics. (For more on this see [Sloman 1985].) The interesting question for defenders of (weak) strong AI, then, is whether this primitive form of semantics provides an adequate basis for the much more sophisticated sort of semantics required for the representation of human mental processes, in which we refer to all sorts of remote objects, postulate unreal relationships, etc.

Further elaboration of these ideas could be used to show that purely internal processes may suffice to support the attribution of intentional descriptions to a computing system even if it is not connected to an 'external' reality via sensors and motors. Of course, in that case the machine's thoughts, hopes, fears, or whatever, would relate only to its own internal states, or to abstract entities such as numbers, formal games like chess, formal languages, etc. This causally disconnected system, if it had a suitable computational architecture (hinted at previously) might even have emotional states, such as the joys and disappointments of mathematical investigations, the excitement of coming close to solving a problem, the fear that a proof might turn out to have a flaw, and so on. But this restriction to thoughts about internal states and abstract entities would not prevent them from being as real as thoughts about the stars, atoms, mice, gravity or the next election.

An argument along these lines would refute the position taken up by a number of Searle's critics, namely that in order to give a machine mental states it is necessary to embed it in causal links with an external world. This is indeed required for certain sorts of mental states, or processes, such as perceiving, thinking about or wanting a *particular* external object, But it is not required for mentality as such, and only something like a behaviourist philosophy of mind could require it. (See Sloman 1986 for more on this.)

Are quite new sorts of virtual machines needed?

It seems that Searle was mostly concerned with refuting any form of AI based on current types of computers.

I believe his refutation of the strong strong AI thesis, though presented obscurely and without deep analysis, is successful, but that his refutation of at least the weakest strong AI thesis has failed, though not for the reasons given by most of his critics.

What remains an open question is whether a really successful AI project requires some totally different sort of computational machine. One possibility is the type of ’connectionist’ machine which is being explored now by a number of AI centres (Feldman and Ballard 1982), such as ’Boltzman machines.’ What Searle has written is not sufficiently precise to indicate whether he thinks he has ruled out the possibility of mental processes in that sort of machine. It is interesting that such machines enhance global reliability (for certain tasks) by extreme redundancy in their representations, whilst undermining reliability by allowing almost anything to interact with anything else, directly or indirectly. This simultaneously provides tremendous learning power, and reduces the predictability of sub-processes. The use of statistical noise in such machines to ensure that they do not settle down at local optima is another example of a tradeoff between local and global reliability. I believe there are many more design principles and trade-offs waiting to be discovered. Perhaps some of them will only be discovered by close examination of faulty arguments aiming to show the whole enterprise is misguided.

Acknowledgement

The work reported here was supported by a fellowship from the GEC Research Laboratories and a grant from the Renaissance Trust. Steve Torrance commented at length on an earlier version, which appeared in the AISB Quarterly, 1985. Others who have made useful comments include Margaret Boden, David Allport, Chris Thornton, Peter Hiett.

References

Monica Croucher *A Computational Approach to Emotions* draft thesis, Sussex University, 1985.

J.A. Feldman and D.H. Ballard, ’Connectionist models and their properties’, in *Cognitive Science 6,* 205-254, 1982.

John Haugeland, (ed), *Mind Design,* MIT Press, 1981.

D. Hofstadter and D.C. Dennett (eds) *The Minds I: fantasies and reflections on self and soul.* Harvester Press 19??

G. Ryle, *The Concept of Mind,* Hutchinson, 1949.

John Searle, ’Minds, Brains, and Programs’, with commentaries by other authors and Searle’s reply, in *The Behavioural and Brain Sciences 3,* 417-457, 1980.

John Searle, *Minds Brains and Science,* BBC Publications, London, 1984.

A. Sloman, *The computer revolution in philosophy,* Harvester Press, 1978.

A. Sloman, and M. Croucher, ’Why robots will have emotions’ in *Proc. 7th International Joint Conference on AI* Vancouver 1981.

A. Sloman, ’What enables a machine to understand?’, in *Proceedings 9th International Joint Conference on Artificial Intelligence,* Los Angeles, 1985.

A. Sloman, ’Reference without causal links’, in L. Steels (ed), *Proceedings ECAI, Brighton 1986*