DAVID COLE

ARTIFICIAL INTELLIGENCE AND PERSONAL IDENTITY

ABSTRACT. Considerations of personal identity bear on John Searle's Chinese Room argument, and on the opposed position that a computer itself could really understand a natural language. In this paper I develop the notion of a *virtual person*, modelled on the concept of virtual machines familiar in computer science. I show how Searle's argument, and J. Maloney's attempt to defend it, fail. I conclude that Searle is correct in holding that no digital machine could understand language, but wrong in holding that artificial minds are impossible: minds and persons are not the same as the machines, biological or electronic, that realize them.

Many workers in cognitive science believe that computers can potentially have genuine mental abilities. John Searle has been a prominent critic of this optimism about the abilities of computers. Searle argues that computers can at best simulate, but not possess, intelligence. If Searle is correct, even though a computer might eventually pass the Turing Test, no computer will ever actually understand natural language or have genuine propositional attitudes, such as beliefs. Searle's argument is interesting both because of its import and because it appears to some to be valid (e.g., Maloney 1987), and to others to be invalid (e.g., many of the "peer" commentators following Searle 1980, Sharvy 1983, Carleton 1984, Rey 1986, Anderson 1987).

The following is a defense of the potential mental abilities of digital computers against Searle's criticism. I shall clarify precisely why his argument is logically invalid. The missing premise that would render the argument valid reflects a form of personal identity theory that Searle may accept but which is widely, and I believe rightly, regarded as false.

However, Searle has indeed, I believe, succeeded in proving that no computer will ever understand English or any other natural language.[1] And he is correct in rejecting the "system reply".[2] But, I shall show how this is consistent with the computer's causing a new entity to exist (a) that is not identical with the computer, but (b) that exists solely in virtue of the machine's computational activity, and (c) that does understand English. That is, showing that the machine itself does not understand does not show that nothing does. We can introduce the concept of a *Virtual Person* (or *Virtual Mind*), an entity that may be realized by the activity of something that is not a person (or mind). Thus, I believe, Searle's argument fails in establishing any limitations on Artificial Intelligence (AI).

Synthese **88**: 399–417, 1991.

Thus, one can show, by a line of reasoning independent of Searle's, that it would always be a mistake to attribute understanding to a computer. The line of argument is inspired by considerations raised by John Locke and his successors (Grice, Quinton, Parfit, Perry and Lewis) in the development of theories of personal identity, a branch of analytical metaphysics perhaps not obviously related to AI. This line of reasoning reveals the abstractness of the entity that understands, and so the irrelevance of the fact that hardware (including the system) itself does not understand. Searle was both right and wrong on this assessment; he wins a battle, but loses the war.

SEARLE'S ARGUMENT

Searle's argument is straightforward. It is possible to write computer programs that produce responses in natural language to questions about some subject domain. Some believe that through such clever programming actual understanding is produced. Others believe that genuine understanding is not *yet* achieved, but it may be in the future with improved programming techniques, larger databases and faster machines. But, as Searle argues, consider that no matter how clever and complex the program, a human could do exactly what the computer does: follow instructions for generating strings of symbols in response to incoming strings.

Suppose, for example, a person (Searle, in the original statement of the argument) who does not know Chinese sits in a room with instructions written *in English* (a "program") that tell one in detail how to manipulate *Chinese* symbols, producing strings in response to the strings given to one. We are to suppose that the instructions are such that they permit successful passage of this variation on a Turing Test: even with trials of indefinite duration those outside the room cannot tell the difference between the room as described and a room that contains a human native speaker of Chinese. Since the instructions tell one what to do entirely on the basis of formal or syntactic features of the strings, without ever mentioning (or revealing) meaning, one can generate Chinese sentences without any understanding of what they mean – indeed without even knowing that they are Chinese sentences. That is, doing exactly what a computer does would not give one the ability to understand Chinese. Therefore, the computer does not understand

Chinese either. Thus, mere programming cannot produce understanding of a natural language.

I wish now to consider three claims made in the course of this argument:

(1) the claim that the person following the English instructions would not understand Chinese;

(2) the inferred claim that a computer following a program would not understand Chinese; and

(3) the inferred final claim in the preceding summary that programming cannot produce understanding of a natural language.

There is some reason to be critical of claim (1). Searle's self-report of incomprehension of Chinese in his scenario conflicts with other evidence, notably the response in Chinese to Chinese questions. One might hold that the person in the room understands Chinese albeit with certain odd deficiencies not characteristic of polyglots: most notably, an inability to translate.[3] One may also be critical of the inference to (2). There are important disanalogies between a human following understood English instructions and a computer running a program – the computer does not literally understand its program "instructions"; the computer would not be conscious of its program; the explicitly syntactic character of the "instructions" in the program is a red herring in that whatever is produced by programming a programmable computer could have been hardwired in a dedicated computer (Cole 1984).[4]

But let us suppose that the crucial premise (1) is true: Searle would not understand Chinese merely by following the instructions in English for syntactic manipulation. Let us also concede (2) for the sake of argument. Nevertheless, (3) does not follow.

Clearly from the fact that *someone* does not understand Chinese it does not follow that *no one* understands Chinese. And from Searle's linguistic disabilities in the Chinese Room scenario, it does not follow that no one *in the room* understands Chinese, unless Searle is *alone* in the room. And, finally and this is the main point here, it does not *follow* logically from the premise, that Searle is initially alone in the room and that no one else *enters* the room from outside, that Searle remains alone in the room.

THE KORNESE ROOM

The Chinese Room argument as it stands is logically invalid. The question then is whether it is legitimate to assume that if anyone understands Chinese, it must be Searle (who does not). It might be thought, since the question-answering performance suggests that there is someone who understands Chinese and that the only one around is Searle, that the only one who could possibly be the one who understands Chinese is just Searle. But it is not necessary that it be the case that the performance of the room suggests that there is any *one* who understands Oriental languages in the room.

To show that, let us consider a variation on Searle's thought experiment. Once again we are to imagine Searle in a room, following instructions for dealing with strings of alien characters slipped under the door. The room is a bit more crowded than before: the size of the set of instruction manuals is almost doubled from before. And the pace is more hectic. Searle flips vigorously through manuals as a steady stream of squiggles and squoggles is slipped under the door. He generates reams of marks in return, as instructed by the manuals. As before, many of those outside believe that someone in the room understands Chinese, for they are receiving replies to questions submitted in Chinese.

But unlike before, those outside *also* believe someone in the room speaks Korean, for they are receiving replies in Korean to questions submitted in Korean. Furthermore, they also believe that the room is more crowded than before: there appear to be at least two people in the room, one who understands Chinese (call him Pc) but not Korean, and another who understands Korean but not Chinese (call him Pk). The evidence seems quite clear that Pc is not Pk. In fact let us suppose that the response abilities and dispositions embodied in the Pc database were derived from an elderly Chinese woman, whereas those in the Pk database were from a young Korean male who was a victim of a truck-bicycle collision.[5] The person, if any, who understands Chinese, is not the person, if any, who understands Korean. The answers to the questions in Chinese appear to reveal a clever, witty, jocular mind, knowledgeable about things in China, but quite ignorant of both the Korean language and events in Korea. Pc reports being seventy-two years old and is both wise and full of interesting observations on what it has been like to be a woman in China for the tumultuous past half-century. By

contrast, the replies to Korean questions reveal quite a dull young man. Pk is misogynous. Pk has a vitriolic hatred of China, but is largely ignorant of events there. Pk is very avaricious, and soon discovers that he can demand money for answering the questions slipped under the door. Pk reports that he works in a television factory and gives accurate descriptions of television assembly. The only other subject about which Pk exhibits much interest or knowledge are the Olympic Games held in Korea.

Thus, suppose that the behavioral evidence is a clear as can be in such a case that there are two *distinct* individuals in the room. Try as they might, interlocutors outside the room can find no hint that there might be but a single person pretending to be two. Information provided in Chinese is unavailable to Pk, even when offers of substantial rewards are made to him in Korean for answers based on the information provided in Chinese.

But behavioral evidence is certainly not decisive here. Indeed behavioral evidence is the very category of evidence called into question in rejecting the adequacy of the Turing Test as a test of mental abilities. Fortunately, a stronger case can be made by considering information not available to the interlocutors outside the room. There is in fact no individual inside the room who understands both Chinese and Korean. Searle understands neither. And the instructions for generating replies to Chinese input and those for dealing with Korean input are distinct, with no exchange of information between the databases consulted (although this is not known by Searle.) If there were a single person duplicitously feigning being two, the person would, on the one hand, realize that something was being asked about events that he/she knew about and then would pretend not to know about them. That never happens in the room. And the histories of the representations in the Pc and Pk databases are completely independent, involving individuals in China and Korea respectively. Thus *if* Pc and Pk are persons, they are distinct.

At this point considerations emerge familiar from arguments in other contexts concerning personal identity (c.f., for example, Perry 1978, pp. 32–36). Pc cannot be identical with Pk. The grounds for saying that Pc is Searle are just the same as those for holding that Pk is identical with Searle. We cannot hold that both are identical with Searle, for this would violate the transitivity of identity. Therefore, we must hold that neither is Searle. Thus, by a line of reasoning quite independent

of that used by Searle, we arrive at the conclusion that the person, if any, who understands Chinese is not Searle.

COMPUTERS AND PERSONS

Consider now a computer system. Suppose that a program has been written for this machine which embodies the very algorithm for replying to questions that was used in the English instructions imagined in the Kornese room scenario. Again, there is no interchange of world-information between the Korean and the Chinese databases. Let us suppose then that the computer system responds to questions asked in Chinese and to questions asked in Korean, with performance indistinguishable from the Kornese room. Now let us suppose that someone wished to say that the computer itself understands Chinese, say, and can answer questions about China. But when the system is asked in *Korean* if it understands *Chinese*, the reply comes back that it does not. Does anyone *lie*? Does *the computer* lie? The behavioral evidence is, *ex hypothesi*, just what it was in the Kornese Room scenario. When asked in Korean about China, the replies do not demonstrate knowledge of China, only a vitriolic prejudice against China. When asked similar questions in Chinese, the replies exhibit knowledge and love of China. Does *the computer* like China or not?

These considerations suggest that it would be a mistake to attribute these properties to the computer itself. One would have equally good grounds for attributing incompatible properties to the computer. For the same reason it would be equally incorrect to attribute knowledge or ignorance of China to the *program*. Again, one cannot attribute inconsistent properties to a single entity. The solution is to hold that no *single* entity understands both Chinese and Korean; there are two subjects, two virtual persons: one who understands Chinese and one who understands Korean.[6] These two virtual subjects are realized by a single substratum, the computer.

The concept of a *virtual machine* is familiar in computer science. And some computer scientists, such as Paul Smolensky, have viewed consciousness as a virtual machine.[7] Each computer has an intrinsic instruction set: the capacity to perform a certain set of operations is wired into the central processor. And in virtue of the construction of the machine, a certain syntactic string will cause one of the intrinsic operations to be performed. But it is possible to write a program which

will cause one machine to behave as a different machine, one with a different intrinsic instruction set. Thus, such emulation software may make a computer built using an Intel 8088 processor behave as though it were a Z80. Then, there is said to be a virtual machine. The virtual machine can run software written for a Z80, using the Z80 instruction set. The 8088 realizes a virtual Z80. Of interest, some newer processors (for example, the Intel 80386) can realize multiple virtual processors concurrently, and this capacity is wired in as an intrinsic capability of the computer.

Now it might be thought that a virtual machine is not a real machine. But there is no reason for this reservation. One machine can very literally realize, or make real, another or several other machines. In fact, a manufacturer could decide to sell very real computers in which the nominal processor was realized by another. For example, a Reduced Instruction Set Computer (RISC) processor might be used, because of its great speed, instead of a normal full instruction set 80386. The user might be quite unaware that his 80386 was in fact a virtual machine, realized on a RISC that was not intrinsically an 80386. And there are physical LISP machines as well as virtual LISP language processing machines.[8] The only difference between the physical and the virtual machine has to do with intrinsic instruction sets, with consequences for speed and volatility. When the emulation software is not running, the virtual machine does not exist.

Note that the physical and the virtual machines differ in properties. They run different programs. At a given time, the physical machine will be running the emulation program, whereas the virtual machine may be running an application. The two machines have different instruction sets. And the speeds at which the two machines perform basic operations will differ. Thus, the physical machine and the virtual machine(s) it realizes are not identical. And the virtual machine is not identical with the emulation program that realizes it when the program is run on the physical machine. The emulation program may be long or short, may be written in a language, may contain comments, may be copyrighted, but the virtual machine has none of these properties.

There are additional considerations that count against holding that the *program* incorporating the Chinese or Kornese Room algorithm understands language or likes China. The program itself is entirely inert until it runs. In Aristotelian terms, a program could be but the form of some matter – without an underlying substance, it does nothing.

The program exists before and after it is run, but understanding, if any, exists only while the program is running.

In the light of this result and consideration of the Kornese room, let us reconsider Searle's original scenario. It is clear that the knowledge, personality, beliefs and desires apparently revealed in the Chinese answers to questions submitted to the room might be quite unlike Searle's own. Indeed, Searle himself could receive no knowledge through information provided in Chinese – and he can reveal none that he has. He cannot express *himself* or anything about himself in the Chinese answers. Similarly, the answers in Chinese reflect no access to Searle's knowledge, preferences, or personality. If there is a person who understands Chinese, it is clearly *not* Searle.

Thus, what follows from the fact that Searle does not understand Chinese is just that the person, if any, who does understand Chinese is *not* Searle. A tacit premise, needed for the inference from (2) to (3), is that there is no *other* mind. But Searle gives us no reason for believing that this premise is true. There may well be a mind *realized* by Searle's activity, a virtual person. But the *same* mind could have been realized by the activity of someone other than Searle – Searle could even resign his job in the Room and be replaced by another – while the Chinese conversation continues. This is additional evidence that the Chinese understanding person is not Searle. Searle is not essential to the existence of the Chinese understanding person.

FUNCTIONALISM AND MULTIPLE MINDS

Georges Rey (1986) advocates a version of the system reply incorporating the robot reply. Rey holds that

> Searle's example burdens [AI] with a quite extreme view about the 'autonomy' of language, a view that would allow that understanding a language need involve *only intra*-linguistic symbol manipulations. (p. 171)

Clearly functionalism will require more integration than that. Rey goes on to consider a robot (a system with sense organs) and argues that, on a sketched in causal theory of meaning, the system would understand language. But Searle is quite clear and correct in pointing out, in his discussion of the robot reply, that his argument is not affected in its essentials by extension to include extra-linguistic capabilities. Therefore, Searle is not (merely) attacking a strawman. Sharvy (1983, p. 128)

also seems too cavalier in saying that Searle's argument against the system reply is just an "intuition pump".

Maloney (1987) disagrees with Rey's rejection of the extreme view of autonomy of language, citing severely handicapped individuals as support for the autonomy. The counterexamples are not conclusive, I believe, because they focus on overt behavior – even the handicapped have the neuronal subsystems that in normal individuals serve motor and perceptual skills. In any case, we can sidestep this issue by simply considering a robot system, for Searle claims that his argument applies equally to the robot system as to the original person confined to the room and to exclusively linguistic input and output.

Cole (1984) and Rey (1986, p. 174–75) hold that Searle, despite protests, understands Chinese in the room. Maloney (1987, p. 355–59) and Cole point to Searle's failure to translate; Cole finds it only odd, but Maloney finds it decisive for denying language comprehension. But then Maloney goes on to consider a view similar to the one advocated here, which agrees that the English-speaking occupant of the room (Maloney calls him "Marco") is not the person who understands Chinese. The view Maloney considers is that "there must be another agent, Polo, who does realize the program and thereby understands Chinese" (p. 359).

But, says Maloney, "there is an overwhelming difficulty with postulating Polo that emerges upon closely examining him" (p. 360). Polo shares Marco's sensory and motor systems and goes wherever Marco goes; thus *how* can Polo be distinct from Marco?

> Marco learned how to manipulate the cards in much the same way in which he has previously learned lots of different things, including poker. According to Strong AI, understanding how to play poker involves mastering the proper program, just as understanding Chinese amounts to running the right program. Now, since Marco both learned how to play poker and also plays poker, i.e. understands poker, why is it Polo rather than Marco who understands Chinese, since it was Marco, not Polo, who mastered the program for Chinese? (p. 363)

Maloney does not wait for an answer:

> all that Marco did in order to understand poker was learn a program. . . . If anyone here understands Chinese, it must be Marco, not Polo. And so, since we have already established that, despite realizing the formal program for Chinese, Marco is ignorant of Chinese, Strong AI is finally and thoroughly false. (p. 363)

This conclusion is premature. Maloney's argument seems most forceful as a critique of Sharvy (1983). Sharvy says:

> First, consider a man who is locked in a room receiving symbolic inputs and calculating symbolic outputs according to a purely formal algorithm, but who does all this completely blind to any interpretation of those symbols. That man cannot truly be said to be playing chess. This is so even if men outside the room – interpret the symbols as representing moves in a chess game. But a computer running that very same program is playing chess and is doing so *by* running that program.
>
> So playing chess is an example of something that computers come to do by instantiating a program, but which a man in a room does not come to do by instantiating that same program. (p. 127)

This position is odd, and surely Sharvy needs an explanation of the sort Maloney demands for the difference. In any case, Sharvy's position is incompatible with functionalist approaches to mentality. My task here is to develop an assessment of the relation of machines to minds which is compatible with functionalism and to show why a functionalist ought to view Searle's argument as unsound.

Maloney's argument does not show, as he seems to think, that there is not a difference between learning to play poker and learning to run the Chinese Room program. For one thing, Maloney appears not to appreciate the possible relations between Marco and Polo. Accordingly, he offers us a false dichotomy in the following passage:

> Strong AI must accept one of two alternatives. Either Marco and Polo are cognitive systems sharing time in the same nervous system, now one using it, now the other, or they must be genuine parallel processors, different programs simultaneously realized in different sections of the central nervous system . . . . (p. 361)

This is not true. When one system realizes another, it is not the same thing as either parallel processing, with both programs running independently and simultaneously, nor time sharing, with now one program using the central processor and then the other, sequentially. The virtual system is realized *now* by the operations *now* of a single implementing system. Both parallel processing and sequential time sharing imply complete independence between the operations of the two systems (except for time delays). For example, if programs A and B are running on parallel processors, the operations performed by A (say, a statistical analysis program) make no difference to the operations of B, which might be a game. The same is true of time sharing. The two programs are logically independent and this is an essential feature of their re-

lation. If they interacted, it would defeat the whole purpose of the system.

The case of a virtual system is quite different. Here the activity of the virtual system A occurs solely in virtue of the activity of the realizing system B. Changes in operations performed in B directly affect the operations in A. And a certain group of B's operations is the same as one of A's operations, hence, the direct effect.

This misunderstanding of the relation between the physical and the virtual system affects Maloney's (and Searle's) failure to see the difference between learning to play poker, for example, and realizing a distinct personality which might understand a foreign language unknown to the physical system itself. The failure is striking in Maloney's case, for in arguing that the physical system ('Marco') does not itself understand Chinese, he presents the considerations relevant to answering the question of how to distinguish learning to play poker from realizing Polo, a Chinese understander.

Not surprisingly, the key difference is psychological integration, that is, access and control. This is not surprising for these are just the characteristics that arise in assessments of multiple personality. When one learns to play poker, one understands the objectives of the game and how the rules constrain the players' pursuit of those objectives. All aspects of play can be affected by other psychological aspects of the player. If one is a risk taker, it will be reflected in one's poker play. And one can explicitly relate poker playing to other games, or to life in general. But this level of psychological integration is precisely what is missing in Marco's activity that produces strings of Chinese characters. None of Marco's psychology is reflected in Polo's performance because Marco has no access to the semantic content of the Chinese characters.

This suggests that we consider two ways of learning to play poker. One would be the usual method, whereby one is told the rules and objectives of the game, as well as various informal strategies for success, and allowed to watch a game or run through a couple of hands. The second would be where someone who did not know how to play poker was given a formal program for manipulating strings of ones and zeros which, unbeknown to one, represented in binary code various combinations of cards and information about the betting behavior of poker players. One could then sit in a "Poker Room", having strings of digits fed to one and, in accord with the program, issue strings in response.

Persons outside the room might interpret these strings as poker play, having been told that they were playing with an eccentric computer scientist recluse.

Would one thereby have learned to play poker? One would deny knowing how to play if asked if one could play. And the actual "play" would in no way reflect one's personality, aversion to risk, flamboyance, memory skills, cunning, or skill in the assessment of the psychology of others. One would derive no more pleasure from 'winning' than from 'losing' – one would not know that one had done any of these things. The same considerations of lack of integration which count against saying that Searle or Marco understands Chinese in the Chinese room count against saying that the physical Poker Room occupant plays poker. Thus, the alleged problematic difference between learning poker and learning to respond to Chinese turns out not to be a difference at all, and so certainly is not a demonstration that Strong AI is "thoroughly and finally false".

FUNCTIONALISM, PERSONS AND BODIES

I do not believe it can be proven that there is a person who understands Chinese in the scenario. But this difficulty is a completely general difficulty familiar as The Problem of Other Minds. My argument so far has been to show that even those who do believe that there might be one who understands Chinese in the scenario should resist any temptation to think, as Searle would have them, that it would have to be Searle. Now I wish to argue that support for the view that it is possible for there to be other than a one-to-one correspondence between living bodies and minds comes from plausible accounts of the relation of mind to body. Then, if the actual replies to questions demonstrate understanding and a personality, it would be an inference to the best explanation that there is a person who understands the questions.

Psychiatry has long recognized cases of multiple personality. The condition is rare (under two hundred total reported cases), but is more frequently diagnosed now than in the past. The American Psychiatric Association (*Diagnostic and Statistical Manual of Mental Disorders, III*) characterizes this disorder as follows:

> The essential feature is the existence within the individual of two or more distinct personalities, each of which is dominant at a particular time. Each personality is a fully integrated and complex unit with unique memories, behavior patterns, and social

relationships that determine the nature of the individual's acts when that personality is dominant. (p. 257)

It is of philosophic interest that the diagnostic indication emphasizes the functionalist feature of integration. The disorder is a subtype of "Dissociative Disorders", a variety that includes retrograde amnesia. The *DSM* reports that usually the 'original' person has no knowledge of any of the "subpersonalities", but that the latter may be aware of one another. Recently the law has had occasion to take note of the phenomenon: William Mulligan was found not guilty by reason of insanity of four rapes. Mulligan displayed ten personalities; "the Lesbian Adelena is thought to be the 'personality' who committed the rapes" (Sarason and Sarason 1987, pp. 138–39). Apparently, some evidence suggests that different areas of the brain are responsible for different personalities (Braun 1984).

In any case, the considerations raised by the Chinese and Kornese Room scenarios mirror familiar metaphysical problems and positions in the philosophy of mind. There appear to be good reasons for holding that it is false to say that I am identical with my body. But this is not to say, with the dualist, that I am identical with some substance other than my body, or identical with a whole composed of two substances. And yet the dualist is correct in holding that I might exist while my (present) body did not.

These, I believe, are metaphysical consequences of functionalism. Functionalists have not generally been concerned with how functionalism bears on traditional metaphysical questions, such as the possibility of immortality, but I believe it has interesting implications for these questions (cf. Cole and Foelber 1984). I shall not defend functionalism here, but shall indicate how it bears on the nature of persons and how this is relevant to Artificial Intelligence and Searle's argument.

Functionalism rejects a type-type identity between psychological states and physical states. Some instances of being in pain, say, may be physiologically different from others. There could even be alien lifeforms with psychological states that were of the *same* type as psychological states had by humans, but that had quite a different underlying physical system (for example, based on silicon rather than carbon).

Furthermore, although this has received less attention, these considerations apply to a single individual across time. My psychological states in 1989 need not be realized by the same physical states as were

my type-identical psychological states in 1979. For example, a portion of my brain may have sustained injury in the interim and its function may have been assumed by a physiologically distinct structure. Or, it may become possible to replace damaged portions of my brain with cultured neonatal tissue that grows to assume functions temporarily lost. Finally, it might even become possible to replace entire damaged neurons by functionally equivalent silicon-based electronic devices.

Contemporary discussions of holistic models of cognitive function also underscore that identity of realizations is not essential for type identity of psychological states over time. If Connectionist or Parallel Distributed Processing models of psychological function are correct, the underlying system that realizes the cognitive states of persons is continually changing as learning takes place. The system is radically dynamic, as each bit of new information slightly changes weightings and probabilities of a given global response.

Functionalism thus takes the underlying substance type to be non-essential to the psychological states. This is not to suppose that there can be psychological states without *any* underlying substratum – the inference from the non-essentiality of any *given* substratum to the non-essentiality of the existence of *some* substratum or other would be a modal scope fallacy. Given that the most reasonable supposition is that dualists are wrong in holding that there is any other than physical substance, and in fact only organic neural substance is capable at present of realizing mental states, the result is clear: no brain, no pain. But this is not to say that in order for *me* to experience pain it must be with *this* brain with each of its *current* constituent cells and molecules.

WHY THE "SYSTEM REPLY" TO SEARLE'S ARGUMENT IS WRONG

The functionalist diachronic perspective on persons suggests that persons or minds are more abstract than a simple identity of a person with a body (or a Cartesian soul or individual *res cogitans*) would suppose. This position is not new in this century; it was an implication of Locke's theory of personal identity which invoked a functional connection, memory, as the glue of the mind. Locke says:

> it must be allowed, that, if the same consciousness . . . can be transferred from one thinking substance to another, it will be possible that two thinking substances may make

but one person. For the same consciousness preserved, whether in the same or different substances, the personal identity is preserved. (*Essay*, Bk. ii, chap. 27)

Locke goes on to indicate that a single substance might be the seat of more than one person (if there is not psychological continuity over time).

This view rejects a simple identity of person with underlying substance, whether that substance is or is not material. Presumably Searle would say that I am identical with my body and could not exist without it, or at least not without the brain. But Locke's view suggests that I could. Another body and brain exactly like this one, with the same "causal powers", could, in principle, replace this one. (Indeed, for all I know this may have happened!)

Note that even in the case of bodies, a simple identification with the physical constituents fails to do justice to the identity conditions we in fact employ: Is my body identical with this particular collection of molecules that now constitutes it? No, my body can change, acquiring and losing constituents. A person is an attribute of a body. A single body might realize more than one person, and a single person might be realized by more than one body.

Thus, the "system reply", as Searle represents it, is not quite right either. The Chinese understanding person is not identical with a whole system composed of Searle, the instruction manuals, and the scraps of paper on which he makes notes. As Searle rightly notes, he could in principle commit the contents of all the instruction manuals to memory and follow the instructions completely in his head. That is, he could sit in an otherwise bare room with only a pen and the pieces of paper upon which he writes the outgoing strings of symbols. Still, he would no more understand Chinese than he did when he consulted the manuals for instruction each step of the way. So, the entity that understands Chinese is not Searle nor Searle and paper and manuals.

It will not do to hold (with Cole 1984 and Rapaport 1990) that Searle understands Chinese but does not understand that he understands Chinese; for the Kornese room scenario shows that contradictory psychological attributes can be had by the (virtual) persons manifested by the system. While a single individual might conceivably understand Chinese but not know this, a single individual cannot plausibly be held in any straightforward sense to both find Chinese music always restful and conducive to thought and also to find this music always to be cacoph-

onous and disturbing. Nor could one believe that Malraux's novels misportrayed events in China and also believe that one had never heard of Malraux. There is no reason to suppose that persons realized in the Chinese room would have psychological properties compatible with one another nor that the realized would have the properties of the realizer. Thus, in the original Chinese Room situation, there is no reason to attribute the psychological properties of the virtual person to Searle himself nor to the system consisting of Searle and inanimate paraphernalia.

So who or what *does* understand Chinese in the Chinese room? An unnamed Chinese person. This person is not Searle, but this person cannot exist unless someone – Searle or any competent other – brings to life the Chinese mind by following the instructions in the room.

CONCLUSION

In the rejection by functionalists of type-type identities of psychological events or states with physical events or states, the way is opened for different persons to have quite different underlying physical realizations of their mental states. Functionalism takes certain of the causal properties of an event to be determinants of the psychological properties. One of the psychological properties thus determined is just which mind the event belongs to. Causality may or may not be the cement of the universe but it is what holds bundles of psychological states and events together to form a single mind over time.

Functionalism does not require a one-to-one correspondence between persons and bodies. Contingently, there generally is such a correspondence, which is exactly what functionalism would lead one to expect. The causal properties of psychological states are just those of the system literally embodying them. And in the ordinary course of events, the physiological characteristics of brains permit psychological integration and continuity for the entire duration of the operating life of the brain. But it *could* be otherwise, and may in fact be. There may be multiple persons embodied by a single brain in cases of "multiple personality". Whether there are or not, I believe, turns on the extent of causal connection (and, hence, access) between the personalities. My impression is that actual cases reported to be of multiple personality typically involve a degree of shared access to information which is not characteristic of distinct persons; the multiple personalities typically

speak the same language and (this is not independent of the shared linguistic abilities) have shared knowledge of general facts. Experience with severed corpus callosum and with various memory deficits, as in Alzheimer's disease, suggests that there may be no well-defined threshold of integration (causal interconnectivity) at which one can say that above this threshold there is a single mind and below it there are two or none. Experience specifically with the split brains demonstrates the contingency of the character and count of minds upon causal features of the underlying system.

From the fact that there is a single physical system, then, nothing follows about the number of minds which the system might realize. Depending on the causal character of the system, it might realize no minds, one mind, or more than one mind. This is the case whether the system employs neurons, as in humans, entire humans, as in the Chinese Room, or programmed computers, as in AI. As a result, Searle's Chinese Room argument shows nothing about the possibilities of artificial intelligence.

NOTES

[1] This represents a rejection of the position I took several years ago in Cole (1984).

[2] I defend Searle against several other criticisms, advanced by Philip Cam, in a paper in the *Australasian Journal of Philosophy* (September 1991). As I reconstruct it here, Searle's reasoning is clearly valid and the indicated conclusions are correct. But, Searle's argument turns on a failure to consider the possibilities of virtual entities discussed here and so fails to refute an interesting version of Strong AI.

[3] See Cole 1984.

[4] A similar point is made in the Searle-Churchlands debate in *Scientific American*. (See Churchland and Churchland 1990 and Searle 1990.)

[5] The derivation should be by whatever causal process that can preserve the representational properties of the brain states of the original biological persons. Xeroxing preserves the representational properties of written information; so does conversion to electronic media. The analogous process for brains might be as detailed as a digitized functional equivalent of the entire brain – a neural net – or (more plausibly) a functional equivalent at a higher level of analysis, such as would be provided by an exhaustive intellectual and personality inventory. However, the former might well be technically simpler and more expeditious, just as xerographic copies are more easily obtained than paraphrases or translations.

[6] As some readers have suggested, one could avoid the incompatible properties problem by attributing understanding to *parts* of the program, with Chinese understanding and Korean understanding attributed to different parts. This tack raises problematic issues concerning the identity conditions of programs, some of which are currently being ex-

plored by the courts. In a nutshell, my response is that programs are abstract, but not as abstract as persons. Since it is a person who understands, and the same person can be realized by distinct programs, the understanding person is not identical with the program. This reasoning parallels the standard functionalist objections to identifying a person with his/her body. These considerations are set out in this paper in the section below on the 'systems reply'. I believe the reasoning was interestingly anticipated by Descartes in his argument for The Real Difference between mind and body, but that topic is beyond the scope of this paper.

[7] See, for example, Smolensky: "We can view the top-level conscious processor of individual people as a *virtual machine – the conscious rule interpreter* – and we can view cutural knowledge as a program that runs on that machine" (1988, p. 4). While I treat distinct persons realized by a single body as virtual persons, Smolensky views a single person as a collection of virtual machines.

[8] My thanks to an anonymous referee for this point. The referee goes on to remark, in support of my main point about the locus of understanding, that "nobody ever suggests that the LISP interpreter 'understands' the application the LISP program encodes".

REFERENCES

American Psychiatric Association: 1987, *Diagnostic and Statistical Manual of Mental Disorders, III*, American Psychiatric Association, Washington D.C.

Anderson, David: 1987, 'Is the Chinese Room the Real Thing?', *Philosophy* **62**, 389–93.

Braun, B. G.: 1984, 'Toward a Theory of Multiple Personality and Other Dissociative Phenomena', *Symposium on Multiple Personality, Psychiatric Clinics of North America*, Saunders, Philadelphia, Vol. 7, pp. 171–93.

Cam, Philip: 1990, 'Searle on Strong AI', *Australasian Journal of Philosophy* **68**, 103–08.

Carleton, Lawrence: 1984, 'Programs, Language Understanding and Searle', *Synthese* **59**, 219–30.

Churchland, Paul M. and Patricia Smith Churchland: 1990, 'Could a Machine Think?', *Scientific American* January, 32–37.

Cole, David: 1984, 'Thought and Thought Experiments', *Philosophical Studies* **45**, 431–44.

Cole, David: 1990, 'Cognitive Inquiry and the Philosophy of Mind', in Cole, Fetzer and Rankin (eds.).

Cole, David: 1991, 'Artificial Minds: Cam on Searle', *Australasian Journal of Philosophy* **69**(3).

Cole, David, James Fetzer, and Terry Rankin (eds.): 1990, *Philosophy, Mind, and Cognitive Inquiry*, Kluwer Academic Publishers, Dordrecht.

Cole, David and Robert Foelber: 1984, 'Contingent Materialism', *Pacific Philosophical Quarterly* **65**, 74–85.

Double, Richard: 1983, 'Searle, Programs and Functionalism', *Nature and System* **5**, 107–14.

Dretske, Fred: 1985, 'Machines and the Mental', Presidential Address delivered before the Central Division of the American Philosophical Association (reprinted in Cole, Fetzer & Rankin (eds.)).

Fields, C. A.: 1984, 'Double on Searle's Chinese Room', *Nature and System* **6**, 51–54.
Grice, H. Paul: 1941, 'Personal Identity', *Mind* **50**, 330–50.
Lewis, David: 1976, 'Survival and Identity', in A. E. Rorty (ed.), *The Identities of Persons*, University of California Press, Berkeley.
Maloney, J. Christopher: 1987, 'The Right Stuff', *Synthese* **70**, 349–72.
Perry, John (ed.): 1975, *Personal Identity*, University of California Press, Berkeley.
Perry, John: 1978, *A Dialogue on Personal Identity and Immortality*, Hackett Publishing Company, Indianapolis.
Quinton, A. M.: 1962, 'The Soul', *Journal of Philosophy* **59**, 393–409 (reprinted in Perry 1975).
Rapaport, William J.: 1986, 'Searle's Experiments with Thought', *Philosophy of Science* **53**, 271–79.
Rapaport, William J.: 1988, 'Syntactic Semantics', in James Fetzer (ed.), *Aspects of Artificial Intelligence*, Kluwer Academic Publishers, Dordrecht.
Rapaport, William J.: 1988, 'Review of John Searle's *Minds, Brains and Science*', *Noûs* **XXII**, 585–609.
Rapaport, William J.: 1990, 'Computer Processes and Virtual Persons: Comments on Cole's "Artificial Intelligence and Personal Identity"', Technical Report 90–13, Department of Computer Science, State University of New York at Buffalo.
Rey, Georges: 1986, 'What's Really Going on in Searle's "Chinese Room"', *Philosophical Studies* **50**, 169–85.
Sarason, Irwin G. and Barbara R. Sarason: 1987, *Abnormal Psychology*, Prentice-Hall, Englewood Cliffs, NJ.
Searle, John: 1980, 'Minds, Brains and Programs', *The Behavioural and Brain Sciences* **3**, 417–57.
Searle, John: 1982, 'The Myth of the Computer', *New York Review of Books*, 29 April 1982, pp. 3–6.
Searle, John: 1984, *Minds, Brains, and Programs*, Harvard University Press, Cambridge MA.
Searle, John: 1990, 'Is the Brain's Mind a Computer Program?', *Scientific American* January, 26–31.
Sharvy, Richard: 1983, 'It Ain't the Meat, It's the Motion', *Inquiry* **26**, 125–31.
Smolensky, Paul: 1988, 'On the Proper Treatment of Connectionism', *Behavioural and Brain Sciences* **11**(1), 1–74 (reprinted in Cole, Fetzer and Rankin (eds.)).
Whitmer, Jeffrey: 1983, 'Intentionality, Artifical Intelligence and the Causal Powers of the Brain', *Auslegung* **10**, 194–210, 214–17.

Dept. of Philosophy
University of Minnesota/Duluth
Duluth, MN 55812–2496
U.S.A.