The Journal of Mind and Behavior
Winter and Spring 2012, Volume 33, Numbers 1 and 2
Pages 27–48
ISSN 0271–0137

Computers, Persons, and the Chinese Room. Part 1: The Human Computer

Ricardo Restrepo

Instituto de Altos Estudios Nacionales

Detractors of Searle's Chinese Room Argument have arrived at a virtual consensus that the mental properties of the Man performing the computations stipulated by the argument are irrelevant to whether computational cognitive science is true. This paper challenges this virtual consensus to argue for the first of the two main theses of the persons reply, namely, that the mental properties of the Man are what matter. It does this by challenging many of the arguments and conceptions put forth by the systems and logical replies to the Chinese Room, either reducing them to absurdity or showing how they lead, on the contrary, to conclusions the persons reply endorses. The paper bases its position on the Chinese Room Argument on additional philosophical considerations, the foundations of the theory of computation, and theoretical and experimental psychology. The paper purports to show how all these dimensions tend to support the proposed thesis of the persons reply.

Keywords: Chinese Room, computation, extended mind

There are few scientific ideas, if any, that have elicited as many responses from diverse fields of expertise as Searle's (1980) Chinese Room Argument. Since its inception, philosophers, psychologists, computer scientists, physicists, and mathematicians have contributed their perspectives on this argument, which was at once a definer of, and a challenge to, the emerging field of cognitive science. Positions on the argument consolidated quickly, where those who saw themselves as cognitive scientists rejected the idea that the mental properties of the Man who figures in the argument are of evidential relevance to the central theses of their field, and virtually all parties have agreed not to question the assumption that the Man performing the computations does not understand Chinese.

Many thanks to Ray Russ and Haruna Suzuki for providing helpful comments in the development of this paper. Correspondence concerning this article should be addressed to Ricardo Restrepo, Ph.D., Escuela de Constitucionalismo y Derecho, Instituto de Altos Estudios Nacionales, Av. Amazonas 37-271 y Villalengua, Quito, Ecuador. Email: ricardo.restrepo@iaen.edu.ec or ricardo.restrepo28@yahoo.co.nz

It is, in my view, time to rethink this two-pronged conventional wisdom. The rejection of these two theses gives rise to the position of the *persons reply*. The first thesis of this reply, and the subject of the present paper, is that the mental properties of the Man appearing in the Chinese Room Argument are evidentially relevant to whether computational cognitive science is correct, and consequently that it is not sustainable to try to debunk the Chinese Room Argument by holding that it is not logically valid. The claim is that the balance of available reasons is on Searle's side on this issue. I will provide positive reasons and new considerations that arise from previously unconnected research, as well as question the arguments of the logical and systems replies for rejecting the Chinese Room Argument. The present proposal, I think, not only provides a better-founded identification of who the pertinent Computer in the Chinese Room is, but also serves to make the question of whether or not the Computer understands, more scientifically tractable. The second thesis of the persons reply picks up where the first leaves off by arguing that the balance of evidence is on the side of the thesis that the Man implementing the program for understanding Chinese does in fact understand Chinese, and appears in a future issue of this journal.

Let me begin with an illustration of the perspective of the persons reply. Suppose a man with a gun wantonly kills the son of the milkman and goes to trial. The man watches his lawyer address the jury thus: "Look, we all agree that it is a man with a gun that performed the actions against the son of the milkman. If this is correct, then it is the system composed of the man *and* the gun which would have murdered the boy. But this in no way implies that my client, sitting here empty-handed before you, is guilty. We should assume innocence before guilt, and the fact that the man pulled the trigger by no means entails that he murdered the boy. In the absence of proof that this man is guilty, I urge you to declare him innocent."

Clearly, the lawyer's argument would not be persuasive and the man, given that he is the entity that killed the boy in the present context, is the murderer. Claiming that he did it with a gun by no means absolves him. The relevant entity is the man, and this is so independently of what weapon he might have used. If the balance of evidence indicates beyond reasonable doubt that the man murdered with a gun, then the jury will be correct to declare him guilty. If the evidence does not show that the man murdered, it is correct to declare him not guilty. That is, the man is the entity that is at the centre of evidential relevance as to whether he is a murderer or not, and it is his guilt or lack of guilt that makes theories about the identity of the murderer true or false. The persons reply suggests similar standards should be applied to the Chinese Room Argument in the following ways: first, it implies that the logical and systems replies are, for reasons similar to the ones we have against the lawyer's, unpersuasive.

Second, just as the man who murders with a gun is the murderer, the Man who computes with the aid of paper, baskets, and pencils ("school-supplies") is the Computer who either understands Chinese, as computational cognitive science predicts, or fails to understand Chinese, thereby providing a refutation of computational cognitive science. A computer is an entity that computes, whether it is a man, a silicon-based entity or otherwise, and it is his understanding of Chinese or lack thereof, that confirms or refutes computational cognitive science. Third, if evidence indicates beyond reasonable doubt that a computer understands, then it is correct to say that it does. And if evidence indicates beyond reasonable doubt that the computer does not understand, then it is correct to say that it does not comprehend.

Given all that has been said about the Chinese Room Argument, many things remain to be clarified and defended about the illustration above — that is part of the subject of this paper. However, it is important to note from the outset that illustrations such as this one cannot be easily dismissed by those who adopt the logical and systems replies. I have frequently found people who have adopted these replies objecting to the illustration because they think that one cannot make an analogy between cases involving criminal acts of a person and his culpability on the one hand, and a case where one person computes and has the mental property in question or not, on the other. If such an analogy could not be made, however, that merely serves to knock down a key form of argument employed by the logical and systems replies (e.g., Block, 2002; Copeland, 2002). This could block significant discussion *ab initio*. However, instead of making a case that such comparisons are somehow unable to be made, I wish to argue, in addition to various other arguments and considerations, that they lead to the opposite results than their proponents think.

Entering the Chinese Room

The goal of Searle's Chinese Room Argument is to refute computational cognitive science. Computational cognitive science can be expressed in the following, somewhat equivalent, ways:

> Strong AI: thinking is merely symbol manipulation . . . . The mind is to the brain as the program is to the hardware. (Searle, 1990, p. 116)

> Computational sufficiency thesis: there is a class of automata such that any implementation of an automaton in that class will have the mental property in question. (Chalmers, 1996a, p. 309)

The differences between these two formulations will not be of interest here. They are formalizations of the theory of computational cognitive science. The strategy Searle designed to refute computational cognitive science consists in the construction of a scenario where the computations supposedly sufficient to understand Chinese get implemented, but the understanding of Chinese does not. The scenario takes various forms. Copeland (2002) calls the version that uses the central scenario the *vanilla argument*. This version consists of a Man who otherwise only speaks English entering a Room where he follows the instructions in a rule-book specifying the rules that guide the manipulation of Chinese symbols characteristically used by genuine Chinese speakers, to participate in a conversation in Chinese with Chinese speakers. In the vanilla argument the symbols are kept in baskets around the Room, and the Man uses the pencil, paper, baskets, and rule-book to perform the computations a genuine speaker of Chinese performs. The Man is behaviorally and computationally equivalent to a speaker who really understands the language. However, it is putatively observed that the Man does not thereby have the mental property in question, namely comprehension of Chinese. Thus, computational cognitive science is false.

To put it syllogistically, the argument is as follows:

1. If computational cognitive science is true, then there is a program P, such that any entity that implements P understands Chinese.
2. The Man implements P.
3. The Man does not understand Chinese.
4. Therefore, the implementation of a program is not sufficient for having certain mental properties.
5. Therefore, computational cognitive science is false.

In this paper we will focus on the scenario used by the vanilla version of the Chinese Room Argument. There are, however, other versions. In the internalized program version, the Man, having memorized the whole rule-book and Chinese symbols, performs the computations inside his head. And in the Chinese nation version, a very large group of people — the population of China — perform the computations, each person doing a very small part of the set of computations at issue. The case made for the two main theses of the persons reply applies more easily, in fact, to the internalized program version than to the vanilla version. I choose to focus on the vanilla version because it is the most paradigmatic and because its intuitive power is stronger than the internalized program version, rendering my case harder to make. Thus, if my argument works for the vanilla version, it likewise works for the internalized program version. The Chinese nation version is a case where no one man can serve as the candidate Computer which might understand Chinese, so it is not telling about a case where one man performs the whole computation, as we will see in detail below.

Is the Man the Computer in the Chinese Room?

A very popular, and one of the original responses to the Chinese Room Argument, is the systems reply (Block, 1980). Here are two articulate and typical expressions of this response:

> There *are* computational states with intentionality — they're the room's. This is the famous *systems reply* to Searle's argument. You in the room are part of another system we can call "you-in-the-room," and it is this other system that has intentional, computational states — states semantically tied to the discourse on Chinese history. Put another way, the relevant system is the *virtual machine* made up of you and your rule book. (Dietrich, 1994, p. 24)

> The best criticisms of the Chinese Room Argument have focused on what Searle — anticipating the challenge — calls the Systems Reply . . . . If the whole system understands Chinese, that should not lead us to expect the [central processing unit] CPU to understand Chinese. (Block, 2002, pp. 71–72, brackets added)

Searle contends that the systems reply is worthless because it "simply begs the question by insisting without argument that the system must understand Chinese" (1980, p. 419). Whether Searle is right about this or not, the logical reply defends computational cognitive science by blocking a more basic and prior step in the argument, in effect, saying that it is irrelevant whether the Man understands or not.

> The flaw in the vanilla argument is simple: the argument is *not logically valid*. (An argument is logically valid if and only if its conclusion is entailed by its premise(s) . . .). The proposition that the formal symbol manipulation carried out by the Man does not enable the Man to understand Chinese . . . by no means entails the different proposition that the formal symbol manipulation carried out by the Man does not enable the System to understand the Chinese story. (Copeland, 2002, p. 110)[1]

In contrast to the systems reply, the logical reply "is a point about entailment. The logical reply involves no claim about the truth — or falsity — of the statement that the [System] can understand Chinese" (Copeland, 2002, p. 111, brackets added). Thus, the logical reply, unlike the systems reply, does not rest its case on supposing that the larger system in the room understands. Rather, it rests its case on the putative fact that computational cognitive science can still be true even if the Man performing the computations does not understand, and consequently that the Chinese Room Argument is ineffective.

Of course, if computational cognitive science itself *implied* that it was a Computer distinct from the Man that understands Chinese, then Searle would have failed to construct a refutation of computational cognitive science, for in

[1] The quote switches Copeland's "Clerk" and "Room" to "the Man" and "System," respectively.

that case the Man would not be the relevant Computer. In other words, the logical and systems replies hold the following opposite thesis: the Man is not the Computer whose mental properties are important to whether computational cognitive science is true or false.

However, Searle (e.g., 1980, 1990, p. 116), the author of the argument, explicitly asks us to consider a scenario whereby the Man is the Computer of the functions characteristic of Chinese by describing a case where the Man implements P. Computational cognitive science hypothesizes that a person understands a language by computing certain functions, the computation of which is sufficient for that understanding, and that any entity that computes those functions will have the mental property in question. Now, if the Man does not understand the target language when he performs those computations, then computational cognitive science is shown to be false. On this count, there is at least a prima facie case that Man is the Computer whose mental properties are relevant to the truth or falsity of computational cognitive science and that the mental properties of any other entity are irrelevant.

Consider again, a murder case. Suppose a fictional character, Ned, held that being someone who performs an unlawful killing is sufficient for being a murderer. In comes John and shows Ned a person who performs an unlawful killing but is not a murderer. To put some flesh in the example, perhaps John shows Ned a man who kills with a gun in a country with unjust laws which punish innocent people who, like the man at issue, are forced to kill in self-defense. Ned looks at the example and admits that the man is innocent of murder in this instance, even if he performed an unlawful killing. However, he counters that this does not matter because, nevertheless, the system composed of the man and the gun is the murderer. So, Ned argues, his theory of what it is to be a murderer, that is, someone who kills unlawfully, can still be true. John says that it is quite a big thing to ask to suppose that the system must be the murderer here. Jack weighs in and says that it does not matter whether the man who performs the unlawful killing in self-defense is a murderer or not in determining whether it is true that it is sufficient for being a murderer that a person perform an unlawful killing. Jack holds this is because whether the man is a murderer or not would not imply that the system is not a murderer.

Ned's and Jack's arguments are what fairytales are made of. Independent of other arguments they could employ, it seems clear, I think, that the arguments are flawed. Whether the person who performs an unlawful killing is a murderer or not is what confirms or refutes the theory that performing an unlawful killing is sufficient for being a murderer. To say that whether the person is a murderer or not is irrelevant is clearly not true. By the standards of Ned and Jack, they would agree with the lawyer in the beginning of this paper that a clear murderer should be absolved because he killed with the use of a gun because he is a mere part of the system composed of the man and the gun. But it seems clear that the relevant entity at issue in these cases is the person involved.

By the same token, persons are not made irrelevant by their computing with the use of school-supplies. If the Man computes with the use of school-supplies, this in no way implies that he is not the Computer whose mental properties are relevant to whether computational cognitive science is true.

There is a generous bank of analogies for this kind of claim. Think of cooks following recipes, climbers climbing rocks, teachers using notes and chalk to teach, among other examples. While cooks are persons who follow externally stored recipes using external ingredients, climbers are persons who use external harnesses and ropes to climb rocks, and teachers are persons who teach external students with external notes and chalk, it is still the case that it is the persons who are the relevant cooks, climbers, and teachers. Similarly, in the Chinese Room Argument it is a person that is the relevant Computer who computes the functions characteristic of Chinese understanding.

It might be inquired whether there are issues specific to our conception of computers and computation which should bar us from affirming that the Man is the relevant Computer. Some may want to claim that commonsense philosophical analogies do not cut it. I have been surprised to find theorists who hold that view about these arguments for the persons reply, but who fearlessly use analogies with companies and criminals in order to support the logical and systems replies. Such theorists cannot legitimately criticize similar uses of these examples for opposite ends because of the cases' commonsense nature, unless they forfeit their position altogether.

There are three important sources of the conception of the computer: the social history, the logico-technological origin, and the modern theory of computation. The various sources have differing degrees of strength in determining the proper understanding of what it is to be a computer. All of them, I think, however, support the view that the Man appearing in the Chinese Room Argument is the Computer important for the truth or falsity of computational cognitive science.

The Social History of the Computer

In the social history of labor, "computer" refers to people with a certain profession. Copeland (2004) writes:

> When Turing wrote "On Computable Numbers," a computer . . . was a human being. A computer . . . was a mathematical assistant who calculated by rote, in accordance with a systematic method. The method was supplied by an overseer prior to the calculation. Many thousands of computers were employed by business, government, and research establishments, doing some of the sorts of calculating work that nowadays is performed by electronic computers. (p. 40)

Thus, people are a prime example of computers. Were managers to have taken advice from proponents of the logical and systems replies they might say to their employed computers: "Look, it was the composite entity of you in the office I supplied, together with the papers, pencils, and erasers that calculated the values for the accounting of the company. Since you are not this entity, I do not owe you a wage." The systems and logical replies apply the same sort of reasoning.

The Logico-Technological Foundations of the Computer

That persons are paradigmatic cases of computers in virtue of the things they do is not a fact confined to historical labor contexts. People are so much the central cases of computers that Turing modeled his notional as well as physical computers on people. Kripke (2006) distinguishes the logical and the computer science orientation to computation, and these two disciplinary foundations lend credence to the idea that persons are central examples of computers. Turing was largely responsible for the establishment of the foundational ideas about modern computers, both on the logical as well as on the physical engineering side. It is in the spirit of making his artificial computing machines more like humans (natural computers), that he notes the following in his logical work:

> We may compare a man in the process of computing a real number to a machine which is only capable of a finite number of conditions q1, q2, . . . , qR which will be called "m-configurations." . . . We have said that the computable numbers are those calculable by finite means . . . . For present I shall only say that the justification lies in the fact that human memory is necessarily limited. (Turing, 1936, p. 59)

Here, Turing explicitly asks us to find symmetries between a person and the machine he is mathematically designing. Relevantly, the construction *limits* a characteristic of the notional computing machine he is designing on the basis of a relevant characteristic of humans. It is clear that Turing thinks of people as the paradigm case of computers, in accordance with which his mathematical design (Turing machines) is constructed.

In the context of dealing with physical computing machines, Turing also models them on people. In the *Programmers' Handbook for Manchester Electronic Computer* he maintains:

> Electronic computers are intended to carry out any rule-of-thumb process which could have been done by a human operator . . . . (Turing, 1950a, p. 1)

Lastly, in the artificial intelligence context he states:

> The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer. The

> human computer is supposed to be following fixed rules . . . . We may suppose that these rules are supplied in a book . . . . He has also an unlimited supply of paper. (Turing, 1950b, p. 444)

Not only are humans explicitly said to be examples of computers, but the fit between the scenario used in the vanilla argument and Turing's description is palpable. Turing's human computer is one who, like the Man in the Chinese Room Argument, follows fixed rules specified in a book and manipulates symbols on pieces of paper. From the point of view of what computers in the workplace are, as well as their logical and engineering-oriented conception, persons are paradigm cases. This lends credence to the idea that the Man thought of in the Chinese Room Argument is the Computer.

Modern Computers Are Entities that Do What the Man Does

Some might argue that Turing's theory is not a good place to look for the theoretical foundations of computation. They may contend that while Turing is a computing pioneer, he is obsolete from the point of view of contemporary understandings of computation. However, Turing's theory lives today. For example, Copeland (1996, p. 335) takes his theory of computation to support the sufficiency of Turing's analysis, and for him "to compute is to execute an algorithm." Algorithms are rules for symbol-manipulation. To execute an algorithm is to do what the rules for symbol-manipulation command. The Man in the vanilla argument is supposed to be doing just that: executing the algorithms specified in the rule-book. If to compute is to execute an algorithm, and a computer is the thing that computes, then it follows that since the Man executes the algorithms, he is the Computer.

Copeland's analysis of computation is:

> Entity *e* is computing function *f* if and only if there exist a labeling scheme L and a formal specification SPEC (of an architecture and an algorithm specific to the architecture that takes arguments of *f* as inputs and delivers values of *f* as outputs) such that (e,L) is an *honest* model of SPEC. (Copeland, 1996, p. 348)

The persons reply proposes that this analysis applies to the Man. The Man is the entity computing the functions characteristic of understanding Chinese (*f*). The Man is the labeled computer whose architecture enables him to execute the formal specification of the algorithms in the rule-book, and to take inputs of Chinese and to deliver Chinese outputs. By hypothesis, the Man is an honest model of SPEC (the formal specification of algorithms and architecture for understanding Chinese), where being an honest model requires that the labeling procedure is performed before the actions of the Man, and that the Man behaves in accordance with the strong conditionals constitutive of the algorithms specified in the rule-book.

There is an alternative suggestion in Restrepo (2009) for being an honest model of SPEC, which eliminates the requirement that honest models have the relevant labeling schemes applied to them before the model performs the actions in question. The reason for shedding this requirement is that SPEC is essentially a theory of certain entities, and theories do not have to be constructed and applied before an entity has the properties posited by the theory. Supposing the theory of general relativity is true, for instance, the world is an honest model of it since before Einstein discovered and applied it. The account agrees with the rest of Copeland's theory of computation. This account also applies to the Man and concludes that the Man is the relevant Computer.

Similarly, appropriate mappings between states of a notional combinatorial state automaton (Chalmers' 1996a and 1996b preferred model) specified in the rule-book and states of the Man can be found. This, according to Chalmers, is sufficient for being such an implementation (Chalmers, 1996a, p. 325). Thus, the foundations of the modern theory of computers support the claim that the Computer relevant to the Chinese Room Argument is the Man, and that the Man is the Computer relevant for the truth or falsity of computational cognitive science.

In the section below, I examine and respond to seven objections that can be raised against the thesis of this paper. I begin with the famous analogy with a corporate entity.

The Corporate Entity Analogy

A popular analogy for people who endorse the logical and systems replies uses corporate entities. As Block puts it, "a company can be guilty of transferring nuclear materials to North Korea even if no individual person in the company is guilty" (2002, p. 71). This may well be true in certain legal senses of "can" and "guilty," which are never defined by Block. Yet, that is, in my view, completely irrelevant because Block's statement would be derived from the fact that one can always concoct a hypothetical scenario wherein the laws are stipulated so that the person or corporate entity can be guilty or not guilty depending on one's desire. I note, however, that those considered guilty of a crime by US foreign policy standards, which I take to be what Block probably has in mind, are never excused of that crime for belonging to some corporate entity, and no one could have legitimately excused bin Laden of the criminal acts he committed because of his corporate memberships. To be fair, neither should one excuse George W. Bush for similar reasons (Bugliosi, 2008). Further, in the vanilla version, the case is even easier to decide, for the relevant corporate entity would have just one member. In a case where a company with just one member is guilty of transferring nuclear materials to North Korea, that person would be guilty. Such a case would be the correct analogy for the vanilla Chinese Room

Argument since there is only one person performing the actions described in that argument. Of course, we could have laws such that when individuals transfer nuclear materials to North Korea in the name of a company they are exempt from guilt. We should note that the moral sense of "guilt" would nevertheless apply and insofar as it is morally wrong to make this transfer, the person would be guilty of it.

Further, even in various cases where the corporate entity has more members than an individual person, individual persons are guilty of crimes. Andy Fastow of Enron is an additional example where an individual is held accountable for corporate crimes. The kind of argument Block employs has the force of Hermann Goering's claim at Nuremberg that it was the Nazi regime that was guilty of crimes and that he was but an innocent cog in the bigger corporate machine. Goering might even have argued that he was not guilty because his actions were in accordance with German law. Goering decides to follow Nazism, and the Man implements the algorithms in the rule-book. The corporate entity analogy is a weak excuse for Goering's performance, and similarly, the Man's putative lack of understanding.

One might object that issues of moral and legal liability do not necessarily track mentality. If one believes this to be the case, then one must think that the abundant arguments of this nature in the logical and systems replies are unpersuasive and do not have any effective force. This of course only supports my case against the logical and systems replies. However, if one is at least willing to take such comparisons as part of the considerations that determine one's position on the Chinese Room debate, then, I think, the case made here should be taken into account.

The Continuous Consolidation Case

Block (2002, p. 70) states that the Chinese Room Argument is "derived" from his Chinese nation argument which appeared in 1978. In the Chinese nation thought experiment, all otherwise monolingual Chinese speakers (the Chinese nation) perform tasks which together result in the performance of computations characteristic of certain mental properties, say, of understanding English. Of course, if all this signifies is that the Chinese Room Argument was inspired by his Chinese nation, I will not object. However, the "derivation" should not say that *because* it is irrelevant that none of the people composing the Chinese nation understand English when they perform the computation, it is irrelevant that a single man implementing the whole program for understanding English does not understand. This would be akin to saying that because hundreds of millions of people cannot fit into the Chinese Room, that one man cannot fit. A more sophisticated version of the derivation might argue the following:

> Premise 1: if it is insufficient for any one person participating in the implementation of the program for understanding English to understand English, when that implementation is composed of n people, then it is insufficient for any one person participating in the implementation of the program for understanding English to understand English, when that implementation is performed by n–1 people.
>
> Premise 2: it is insufficient for any one person participating in the implementation of the program for understanding English to understand English, when that implementation is performed by 1.3 billion people.
>
> Line 3: therefore, by Premise 1 and Premise 2, it is insufficient for any one person participating in the implementation of the program for understanding English to understand English, when that implementation is performed by 1,299,999,999 people.
>
> . . . (that is, and so on, iteratively applying Premise 1 to Line 3 and subsequent results)
>
> Line 1,300,000,001: therefore, by Premise 1 and Line 1,300,000,000, it is insufficient for any one person participating in the implementation of the program for understanding English to understand English, when that implementation is performed by one person.

Chalmers (1996b, pp. 324–325) seems to endorse a version of this argument when he imagines how millions of demons interact to implement the functional organization of the brain while none of them understand English. The demons start to double up on their work, while they proportionately diminish in number. The end of the series is a single demon doing everything the brain does and who does not understand English. The argument, however, I think is not successful. It is an instance of the sorites fallacy. Consider the case of baldness: if a man with n hairs is not bald, then a man with n–1 hairs is not bald. One hair, it is believed, cannot make the difference between not being bald and being bald. But if one applies this principle repetitively, beginning with a man who is clearly not bald, one ends up with a man with no hair and the obligation to declare him not bald, which is absurd.

First, it is important to be clear that we are ill-advised to apply this form of reasoning because we know from the beginning that it leads to evidently wrong conclusions. Second, there is an enlightening reconstruction of the sorites reasoning using fuzzy logic which does not lead us to erroneous results. Sorites-type arguments use a premise of the general form:

If x with n y's is Q (or not Q), then x with n–1 y's is Q (or not Q)

Suppose this premise is not entirely true; it has a degree of truth-content somewhere between complete falsity (0) and complete truth (1). The supposition that it is completely true leads to absurdity. The degree of falsity is practically negligible for local applications. However, with each application of the partially false premise, the degree of falsity is accreted onto derived conclusions. Eventually, the degree of falsity becomes more dominant than the degree of truth, and eventually the derived conclusions are completely false. The conclusion that "a man with zero hairs is not bald" was an example. Fuzzy logic correctly enables us to declare this statement false (Copeland, 1997). Thus, if Premise 1 is mostly, but not completely true, as is the analogous premise in the case of baldness, then as one applies it over and over, consolidating the implementation until it is performed by just one man, the conclusion that this person does not understand English is false.

The Personal Computer and the Hat

Suppose someone theorized, as Block does, that the Man is but a central processing unit and that the baskets hold the memories which are a constitutive part of the mind of the implemented computer. And suppose that person concludes, as a result, that the mental properties of the Man are irrelevant to the mental properties of the mind to which those memories belong. In the context of the Chinese Room Argument debate, those who agree with Block typically hold that persons are computers. Consequently, these people should believe that the Man is a computer. This is not implausible since the Man himself holds memories in his brain, so it would not be the case that all of his memories would be located outside his head if he was the Computer to whom the Chinese program and memories belonged. Taking this into account, it seems better to say that rather than that the Man is a mere central processing unit of some other computer, instead what happens in the vanilla argument is that the Man is a computer who stores information about the Chinese language in the Room. The Computer–Man "connects up" to the information in the Room when he starts executing the algorithms specified in the rule-book. The information in the Room is like the information in an external hard-drive.

When one connects an external memory hard-drive to a von Neumann personal computer, one does not get a *new* and distinct computer. There is a change in the computer when one connects it to an external hard-drive. However, the computer survives this change and is the computer to which the information in the hard-drive belongs while it is connected. Similarly, there is a change in the Man when he connects up to the information in the Room. However, he keeps his identity as the relevant Computer to whom that information belongs while he is in charge.

This case can be additionally strengthened by considering a possible scenario whereby the part of the brain that sustains the English language in the Man is extracted from his skull. Such a part of the brain would include Broca's and Wernicke's areas. The procedure can be idealized so that the relevant bits are concentrated there, the extraction does not harm any other part of the Man, and the connections that the English Language Device has to other parts of the brain remain functionally intact. We can suppose the Man in that scenario wears a meter-high hat holding those bits of his brain. The longer distance the signals need to travel is proportionately compensated with axons of super-fast transmission capacity. The Man is now in a similar situation to the one in which he is in the vanilla argument. The school supplies are like parts of his brain he stores externally. In the Chinese Room, as in the case where he holds parts of his brain out of his skull, it is still the Man that speaks and understands the target language — not some other system of which he is a mere part.

Now, if one were to argue that it is irrelevant whether or not the Man understands English because the Computer is larger than the Man (it is Man + external brains), she or he, I think, would be wrong. It is the Man that is the relevant entity that either understands or fails to understand.

The Significance of "Could a Computer Think?"

It might be suggested that what has been said up until now trivializes the question of whether a computer can think, since by the proposed standards, humans, a prime example of thinkers, are a prime example of computers; so it trivially follows that computers can think. However, this should be looked upon as a bonus, as opposed to a difficulty for the persons reply. The question only looks trivial because we now have a decisive affirmative answer. The more principled solutions to questions we obtain from a theory, the better that theory is. Theories are designed precisely for such purposes. There are, however, core related questions which remain open for further research. One is whether there are some computations the implementations of which are sufficient to generate the instantiation of mental properties. The fact that we are computers with mental properties does not prove that having certain computational properties is sufficient for having certain mental properties. Further, even the fact that the Man understands does not entail that the symbol manipulation is sufficient. It is, after all, a Man with independent mentality who is performing the computations. Perhaps an entity that does not have independent mentality would not understand Chinese when it implements the program that the Man implements.

Another question left open is whether *we* think in virtue of implementing certain computational properties, rather than some other mechanisms. We might be computers with mental properties, but we possess those mental properties in virtue of some computation-irrelevant causal powers. An additional issue is whether there are alternative grounds for realizing mental properties. Even if we have mental properties in virtue of the computational properties we possess, could mental properties be realized through computation-irrelevant means? Another question is whether or not humans will be able to artificially build, in a narrow sense, a computing thinking machine. The truism that some machines, and in particular, some computers, think does not definitively settle these questions.

The Extended Mind Suggestion

It might be thought that the persons reply implies a commitment to extended mind theory. Extended mind theory claims that cognitive processes may take place outside a person's skull (Clark, 2006; Clark and Chalmers, 1998). Consequently, one might think that the Man is a mere part of the relevant Computer-mind. However, extended mind theory does not imply that the Man is not the Computer whose mental properties matter. Consider Clark and Chalmers' parity principle:

> If, as we confront some task, part of the world functions as a process, which were it to go on in the head, we would have no hesitation in accepting as a cognitive process, then that part is for that time part of the cognitive process. (cited by Clarke, 2006, p. 44; originally in Clark and Chalmers, 1998)

Now, those of us who say that were the Man's symbol manipulations to take place in his head they would be cognitive processes, according to the parity principle, should also say that the Man's external symbol manipulations are part of the cognitive processes. However, these would still be cognitive processes of the Man, not necessarily of some other entity. Clark and Chalmers write:

> Now consider Otto. Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory. Today, Otto hears about the exhibition at the Museum of Modern Art, and decides to go see it. He consults the notebook, which says that the museum is on 53rd Street, so he walks to 53rd Street and goes into the museum.
>
> Clearly, Otto walked to 53rd Street because he wanted to go to the museum and he believed the museum was on 53rd Street. And just as [a normal person has] her belief even before she consulted her memory, it seems reasonable to say that Otto believed the museum was on 53rd Street even before consulting his notebook. For in relevant respects the cases are entirely analogous: the notebook plays for Otto the same role that memory plays for [a normal person]. The information in the notebook functions just like the information constituting an ordinary non-occurrent belief; it just happens that this information lies beyond the skin. (Clark and Chalmers, 1998, p. 11, brackets added)

It would seem, then, that extended mind theory does not imply that if a person has some externally stored belief, that the belief should be attributed to some entity other than the original agent. For it is Otto, the original person, that has these externally stored beliefs. This is why extended mind theorists say that the environment would be an external memory bank, that is, external to the skull of the person the beliefs are attributed to, but nevertheless the person's. According to extended mind theory, the person might well extend beyond the boundaries of the skull, just as the mind does. Thus, whether this is so or not, application of extended mind theory to the vanilla Chinese Room Argument does not imply that an entity other than the Man has the mental properties relevant to the truth of computational cognitive science.

The Multiple Personality Disorder Hypothesis

It is sometimes thought that the Man in the Chinese Room, when asked in English, would not be able to respond in English and would need to give answers which do not cohere with one another (Block, 1995). Further, Block presumes that what explains the supposedly conflicting answers, or abilities to answer, is that there must be two people involved and that consequently it is somehow admissible that the Man does not understand Chinese.

However, a less contrived alternative is that the Chinese and English responses are normally consistent and the Man knows English, as well as Chinese. There is nothing in the Chinese Room thought experiment that logically implies that the Man's answers are not consistent or that he must understand only one of the two languages. Block (1995, p. 419), however, supports the claim that the responses are inconsistent or correspond to incompatible abilities, with an analogy with a 9-to-5 job. While at work, the Man performs the tasks for understanding Chinese and does not speak English. Out of work, he does not speak Chinese. With this, Block proposes to explanatorily excuse the Man of not understanding English while he is implementing the program while at work, not understanding Chinese outside of work, and the Man consequently giving answers at these two times that are incoherent with one another.

The supposed inconsistency in the abilities and answers of the Man is completely unpersuasive. Block's argument asks us to grant the idea that the Man does not understand English when he speaks Chinese and vice versa because that is what would happen if he had a job where he was required to speak only one of his languages during office hours. But this is surely not what would happen in the vast majority of cases, and certainly bilingual people with jobs do not necessarily have multiple personality disorder. Further, it is a precondition of the Chinese Room Argument scenario that the Man understand English while he performs the Chinese computations. For this is the language in which the rule-book is written and the understanding of which enables the Man to behave just like a genuine speaker of Chinese in the first place. Consequently, when the Man simulates Chinese, he must speak English. So, contrary to Block's grounds for saying that the Man has multiple personality disorder when the Man appears to speak Chinese, the Man also understands English when he speaks Chinese.

Bringing in multiple personality disorder seems to me to be merely a way of muddling the issues so that a desired conclusion can appear to be derived by those who identify themselves as cognitive scientists. Decisions on the personal identity of people with multiple personality disorder are characteristically tricky (Humphrey and Dennett, 1989). Nevertheless, suppose for a moment, for argument's sake, that there is a largely disintegrated set of behaviors exhibited by the Man, which warrant the attribution of multiple personality disorder to him.

Neither of the two interpretations of multiple personality disorder makes Block's claims effective.

One interpretation of multiple personality disorder is that the multiple personalities belong to the same person and that they work somewhat similarly to how memory functions. Memories belong to the person who has them — they are sometimes accessed and sometimes not, and sometimes they interact at a time. Nevertheless, there is just one person, with multiple memories accessed at various times. Similarly, multiple personalities might belong to the person who has them and the different personalities take control at various times and sometimes interact at a time. Under this supposition, multiple personality disorder does not imply that the mental properties of the Man do not matter. It would matter for the truth of computational cognitive science whether the Man understands Chinese through one of his personalities.

The other interpretation of multiple personality disorder implies that there are at least two fully distinct people in one body: one monolingual (the Man) and another who also speaks Chinese. By Block's reasoning, the Man would not in fact understand Chinese; some other person would. However, the problem for this position now becomes that the Man is no longer the implementation of the Chinese program, since it is some other person who has taken control and is performing the actions dictated by the rule-book. Under this interpretation, the Man is not the person performing the computation. Someone else is. Consequently the Man's mental properties do become irrelevant. However, this means that the original thought experiment is no longer being considered, so the response fails to address a scenario in which the putative implementation of the program for understanding Chinese does not understand. While this is true, it still remains relevant for computational cognitive science whether that other person who implements the program does understand, and Block agrees that this other person does understand.

The Child and the Calculator

If a fourth grader gets a 100% on a multiplication test by illegally using a calculator, does he understand multiplication? Surely, there are some such children who do not understand multiplication. This kind of case is indeed troubling for the persons reply. However, I make three points. First, the flip-side of this case is the murdering-with-a-gun scenario, whereby we are disposed to say that nevertheless it is the man who murdered. So we seem to be in a situation whereby on this specific issue, both the systems and logical replies on the one hand, and the persons reply on the other, have equal gains and equal losses in terms of cohering with our intuitions. Secondly, the persons reply could say that the children with calculators during the examination do understand, while admitting that this is not enough to give them a good grade since the test looks to grade based upon more long-term biological understanding. The systems reply also has an epistemic cost here, since it is committed to saying that the system of the child and the calculator do understand multiplication; and the logical reply has the epistemic cost of saying that it does not matter whether the child understands or not. The third point is that commonsense comparisons, like with cooks, teachers, and climbers, as well as the three fundamental foundations of computation, support the persons reply over the others. The balance is that the persons reply is more robustly evidentially supported than the logical and systems reply.

A Methodological Coin

In an important cut of experimental methodologies, psychologists form a hypothesis about a psychological property of humans; they go out and get a sample of humans, put them under various experimental conditions with which the sampled humans interact, measure the variables in which they are interested, and analyze the results to see whether the set of data confirms or refutes the theory about the psychological properties of persons. This elementary methodology assumes that when these kinds of hypotheses are formed, the entity referred to is the person, and that when the set of data comes in, it says something about the person's psychology. The theories originally formulated and later confirmed are not about a "system" other than the person partaking in the experiment. The fact that participants interact with pencils, papers, notes, strings, pictures, glasses, other computers, and other parts of the experimental set-up does not change that fact. The considered methodological coin has the following two sides:

> Hypothesis identification side: psychological theories *to be* experimentally tested on humans are hypotheses about the psychology of persons.
>
> Experimental testing side: psychological theories experimentally *tested* on humans are confirmed or disconfirmed by the results of the tests.

From the perspective of experimental psychology, the Man is a participant in an experiment designed to test computational cognitive science. The experimental set-up is as Searle describes it, and it is designed to test whether the Man understands Chinese as computational cognitive science predicts. Realism in this domain involves consistently upholding the identified elementary methodological assumptions.

Now consider the logical and systems replies. As Block (2002, p. 72) puts it, the logical reply says that "the system may understand Chinese even if no person who is part of the implementation understands Chinese" and the systems reply adds that "the whole system — man + program + board + paper + input and output doors — does understand Chinese." In order to make their respective cases, the logical and systems replies must deny that computational cognitive science is a hypothesis that can be applied to the Man and that the results obtained from measuring the mental properties of the Man support or disconfirm computational cognitive science.

The rejection of the identified methodological principles allows the logical and systems replies to say what they say about the scenario in the Chinese Room Argument. However, applying the standards of the logical and systems replies serve to make the theories about the psychology of persons experimentally unverifiable. Behavioral experiments of the kind considered here always involve an interaction between a person and the experimental environment. If these standards were accepted, Searle's claim that computational cognitive science is not an empirical theory would need to be conceded (Searle, 1990, p. 120; 1992, p. 225; Restrepo, 2009). For any theory tested on a human and any result obtained, one could argue that it does not matter what the mental properties of the participants are and that the theory is true or false of systems of which participants are mere parts.

Consider, for concreteness, Gernsbarcher, Varner, and Faust's (1990) test of the structure building framework theory. The structure building framework theory of comprehension implies that human comprehension works by first using initial incoming information to form foundations (structures) about the general topic a person aims to understand. Then, if incoming information (substructures) is coherent with the general foundation, it is mapped onto those general foundations. If the incoming information is not coherent with the foundations, then a new foundation on which new information can be mapped is formed. Structure building framework theory predicts that poor comprehenders tend to develop too many unconnected substructures without general integrating foundations. Thus, Gernsbarcher, Varner, and Faust expected that good comprehenders perform worse at comprehension tasks when the ordering of a presented narrative is scrambled than when it is not, and that poor comprehenders perform similarly when the narrative is quite coherent and when it is scrambled. The experimental set-up involves interactions by the persons with computers, monitors, and projectors.

Gernsbarcher, Varner, and Faust (1990) found structure building framework theory to be confirmed by the results of their test. Now if the standards of the logical and systems replies are upheld — namely, rejection of the identified methodological coin — possible critics of the experimenters' research would say that (a) these experiments indicate nothing about whether persons comprehend by using mental processes described by structure building framework theory because the persons were interacting with external elements pertaining to the experimental set-up; and (b) in any event, the psychology of the person is irrelevant to whether the larger system understands by using a structure building framework. Independent of whether structure building framework theory is true, asserting (a) and (b) would surely be odd moves. The fact that no psychologist has adopted them is additionally indicative of their oddity.

But now consider the possibility that the researchers' experiments found negative results for their theory, putatively lending epistemic weight to the idea that persons do not use the structure building framework to comprehend. Similarly, someone could argue again that, to the contrary, (a) the experiments indicate nothing about the mental properties of the persons so the structure building framework is neither confirmed nor disconfirmed, and that (b) in any event, the psychology of the person is irrelevant to the truth or falsity of the structure building framework. The moral is that once we open up the possibility of a systems/logical-type response, psychological evidence will always be susceptible to deflection. The rejection of the methodological coin by the logical and systems replies leads to an unattractive unverifiability result for not only computational cognitive science, but also for experimental psychology. The persons reply, instead, gets an additional relative confirmational boost by not having this implausible consequence. Instead, it is in harmony with experimental psychology and, having placed the Man in an experimental context, it provides a way of thinking about cognitive science in a manner that makes its theses more empirically decidable.

In this paper, I suggested reasons for thinking that the Computer in the Chinese Room is the Man, and that consequently, it matters to the truth or falsity of computational cognitive science whether he understands Chinese. If this proposal is plausible, attention should now focus on discerning whether the Man–Computer understands Chinese when he implements the program that characterizes having this mental property. The consensus view that he does not should not be assumed without argument — it needs to be proved.

References

Block, N. (1978). Troubles with functionalism. In W. Savage (Ed.), *Perception and cognition: Issues in the foundations of psychology, Minnesota Studies in the Philosophy of Science,* 9 (pp. 261–325). Minneapolis: University of Minnesota Press.

Block, N. (1980). What intuitions about homunculi don't show. *Behavioral and Brain Sciences, 3,* 425–426.

Block, N. (1995). The mind as the software of the brain. In D. Osherson and E. Smith (Eds.), *Thinking: An invitation to cognitive science, 3* (second edition, pp. 377–426). Cambridge, Massachusetts: MIT Press.

Block, N. (2002). Searle's arguments against cognitive science. In J. Preston and M. Bishop (Eds.), *Views into the Chinese Room: New essays on Searle and artificial intelligence* (pp. 70–79). Oxford: Oxford University Press.

Bugliosi, V. (2008). *The prosecution of George W. Bush for murder*. New York: Vanguard Press.

Chalmers, D. (1996a). Does a rock implement every finite-state automaton? *Synthese, 108,* 309–333.

Chalmers, D. (1996b). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Clark, A. (2006). Memento's revenge: The extended mind, extended. In R. Menary (Ed.), *The extended mind* (pp. 43–66). Cambridge, Massachusetts: MIT Press.

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis, 58*, 7–19. (Retrieved on Aug. 25, 2010, from http://consc.net/papers/extended.html)

Copeland, J. (1996). What is computation? *Synthese, 108*, 335–359.

Copeland, J. (1997). Fuzzy logic and vague identity. *Journal of Philosophy, 94,* 514–534.

Copeland, J. (2002). The Chinese room from a logical point of view. In J. Preston and M. Bishop (Eds.), *Views into the Chinese Room: New essays on Searle and artificial intelligence* (pp. 109–123). Oxford: Oxford University Press.

Copeland, J. (2004). Computable numbers: A guide. In J. Copeland (Ed.), *The essential Turing* (pp. 5–57). Oxford: Oxford University Press.

Dietrich, E. (1994). Thinking computers and the problem of intentionality. In E. Dietrich (Ed.), *Thinking computers and virtual persons* (pp. 3–35). San Diego: Academic Press.

Gernsbarcher, M.A., Varner, K.R., and Faust, M.E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 430–445.

Humphrey, N., and Dennett, D. (1989). Speaking for ourselves: An assessment of multiple personality disorder. *Raritan, 9,* 68–98.

Kripke, S. (2006, June). From Church's thesis to the first order algorithm theorem. Paper presented at the Conference on the Origins and Nature of Computation at Tel Aviv University, Tel Aviv.

Restrepo, R. (2009). Russell's structuralism and the supposed death of computational cognitive science. *Minds and Machines, 19*, 181–197.

Restrepo, R. (forthcoming) Computers, persons, and the Chinese Room. Part 2: The man who understood. *Journal of Mind and Behavior*.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3,* 417–457.

Searle, J. (1990). Is the brain's mind a computer program? In T. Schick Jr. and L. Vaughn (Eds.), *Doing philosophy: An introduction through thought experiments* (pp. 115–121). Mountain View, California: Mayfield Publishing. [Reprinted from *Scientific American, 262*, 26–31]

Searle, J. (1992). *The rediscovery of the mind*. Cambridge, Massachusetts: MIT Press.

Turing, A. (1936). On computable numbers, with an application to the *Entscheidungsproblem*. In J. Copeland (Ed.), *The essential Turing* (pp. 58–96). Oxford: Oxford University Press. [Reprinted from Proceedings of the London Mathematical Society, 42, 230–265]

Turing, A. (1950a). *Programmers' handbook for Manchester Electronic Computer*. Royal Society Computing Machine Laboratory. Manchester: University of Manchester.

Turing, A. (1950b). Computing machinery and intelligence. In J. Copeland (Ed.), *The essential Turing* (pp. 441–464). Oxford: Oxford University Press. [Reprinted from *Mind, 59*, 433–460]