The Journal of Mind and Behavior
Summer and Autumn 2012, Volume 33, Numbers 3 and 4
Pages 123–140
ISSN 0271–0137

Computers, Persons, and the Chinese Room. Part 2: The Man Who Understood

Ricardo Restrepo

Instituto de Altos Estudios Nacionales

This paper is a follow-up of the first part of the persons reply to the Chinese Room Argument. The first part claims that the mental properties of the person appearing in that argument are what matter to whether computational cognitive science is true. This paper tries to discern what those mental properties are by applying a series of hypothetical psychological and strengthened Turing tests to the person, and argues that the results support the thesis that the Man performing the computations characteristic of understanding Chinese actually understands Chinese. The supposition that the Man does not understand Chinese has gone virtually unquestioned in this foundational debate. The persons reply acknowledges the intuitive power behind that supposition, but knows that brute intuitions are not epistemically sacrosanct. Like many intuitions humans have had, and later deposed, this intuition does not withstand experimental scrutiny. The second part of the persons reply consequently holds that computational cognitive science is confirmed by the Chinese Room thought experiment.

Keywords: Chinese Room, psycholinguistics, Turing test

The debate about the Chinese Room Argument is one of the most prominent lines of inquiry for computational cognitive science. Positions around this argument consolidated quickly since Searle introduced it in his classic 1980 "Minds, Brains, and Programs," and it might now appear that everything has been said for and against the argument, leaving interested parties with the task of merely choosing from the available positions. Earlier, I provided, however, new arguments against a central thesis of key responses to Searle's thought experiment, from which one of the two central theses of the persons reply emerges (Restrepo, 2012a). By depicting a person implementing a program for understanding Chinese who "undoubtedly" does not understand Chinese, Searle aimed to show that the

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

theory of computational cognitive science is false. Computational cognitive science theorises that some computations by themselves are sufficient for having certain mental properties. So if Searle is right that the Man implements the computations at issue but does not have the mental properties in question, then computational cognitive science cannot be correct.[1] Searle's main detractors have all claimed that the mental properties of the Man who figures in the argument are irrelevant to whether computational cognitive science is true or false. The first central thesis of the persons reply is that this is not the case. It sustains that the key arguments of Searle's main detractors have significant holes, that there are sufficient reasons emanating from our conceptions of what a computer is to think that the Man is the Computer whose mental properties matter to whether computational cognitive science is true or false, and that to deny this renders an important portion of psychological theories unverifiable. In this regard, the persons reply provides a new way of siding with Searle's much-questioned stance on this issue. However, neither Searle nor virtually any of his detractors question that the Man does not in fact understand Chinese when he implements the program for understanding it. But if the first thesis of the persons reply is correct and the Man does not understand Chinese, then computational cognitive science must indeed be false. Abelson (1980) was the notable exception to having accepted the view that the Man does not understand Chinese. The thought Abelson had, however, has been largely ignored and has remained far from fully developed, explored, and justified. This paper aims to fill many of these gaps in order to develop the second thesis of the persons reply.

The persons reply takes Searle's claim to heart that cognitive scientists should uphold scientific realist standards and be "interested in the fact of . . . mental states, not in the external appearance" (2002, p. 61). With this focus in mind, the Man in the Chinese Room is taken to be a participant in an experiment designed to test computational cognitive science, and the external appearances are taken as mere symptoms of their inner causes. Experimentally structured appearances are the empirical basis of scientific knowledge, and the present paper aims to bring this basis to judgments on the supposed understanding or lack of understanding of the Man in the Chinese Room. If the Man provides robust experimental evidence that he understands Chinese, then we can say that he understands Chinese and computational cognitive science is confirmed. If the Man, on balance, displays evidence that he does not understand, then it would be correct to say that the Man does not understand, and that computational cognitive science is false. The result of applying this elementary scientific realist method, I will try to demonstrate, is the second thesis of the persons reply: that the Man mentioned in the Chinese Room Argument understands Chinese when he implements the

[1]Formalizations of both the Chinese Room Argument and computational cognitive science are found in Restrepo (2009, 2012a).

program for understanding Chinese, and consequently, that computational cognitive science receives confirmation. I must note from the beginning that not all evidence for or against a scientific theory is of equal strength, and that there may be consequences of a correct theory which are not at all intuitive. The proposed theory is no exception. However, what matters in theory choice is that the balance of evidence is distributed more heavily toward one theory rather than the others. Evidence is the data of how things seem to be. Theory choice is determined by putting all the available data together and seeing what, on balance, they are more likely to be an appearance of. In the following discussion, there will be evidence of varying degrees of strength, and no one piece of data is definitive. The cumulative effect, however, is robust support for the persons reply.

In a certain sense, we know from the outset what the result of a battery of tests applied to the Man will be. After all, it is known by hypothesis that the Man is computationally and behaviorally identical with a genuine Chinese speaker, so he will, in tests, perform indistinguishably from such a genuine speaker of Chinese. However, I believe applying the tests reveals details of the implication of this supposition which are otherwise obscured. Applying these tests shows how contrary to scientific realist expectations it is to suppose that the Man behaves as he does under the experimental set-up, while not understanding the target language. The tasks are not trivial and the way in which the Man performs, quite plausibly requires picking up the semantics of the Chinese symbols. Like in a psychological experiment operating with random participants to detect their psychology, the proposal is to take the Man to be a random participant in an experimental environment designed to see what can be learned about his linguistic psychology.

It should be noted that implementing the experimental set-up immanent in the Chinese Room Argument is practically impossible. The Man would have to be moving around the Room, reading and writing much faster than a normal human ever could, in order for his computational actions and deliverances to be indistinguishable from those of a genuine speaker of Chinese. Estimates of the computational power of the brain (in this case, one that understands Chinese), would indicate this, since they range in the astronomical.[2] Nevertheless, we can ask, what would the application of a battery of tests designed to test the

[2]For an index of estimates of computational powers of brains see Sandberg and Bostrom (2008), Appendix A. For illustration: Dix (2005) calculates that a brain with 10 billion neurons could perform 10^{16} computational operations per second (given many assumptions). Perhaps the operations of the Man, corresponding to his Chinese simulation activity, would comparably correspond to the work performed by 10 billion neurons in the brain of a Chinese speaker (a fraction of his total number of neurons). Supposing the Man used all his capacity and each operation (conservatively) corresponded to one centimetre in the movement of his body, he would move 10^{11} kilometres each second, which is significantly faster than the speed of light.

theory that the Man understands Chinese tell us about the psychology of the Man? We can perform experiments on the Man as we would on any other person to ascertain the properties of her psychology. The hypothesis proposed here is that the performance of the Man in the envisaged experimental set-up would lend evidence to the theory that he does what a human does in virtue of which she understands.

I would like to make available to the ensuing discussion a version of the Chinese Room Argument in which the relevant languages are switched. Suppose, instead of the Man being an otherwise monolingual English speaker, that the Man is an otherwise monolingual Chinese speaker. The Chinese Man enters the Room, which contains baskets of English symbols and a rule-book, which the Man uses to manipulate the symbols in accordance with the program for understanding English. The resulting version of the Chinese Room Argument preserves all its relevant logical features. Call this resulting argument the *English Room Argument*. The same relevant question can be put forth. Does the Chinese Man understand English? If the Man in the English Room Argument understands English, then the Man in the Chinese Room Argument understands Chinese. The two are in exactly analogous positions.

One reason I make this version available is for expositional ease. The experimental research in psychology to be applied was conducted in English, and readers of this present paper speak English, while infrequently speaking Chinese, so it makes sense to use the common symbols we ourselves recognize. Perhaps more importantly, I make this version available because it might help deconstruct misleading presuppositions I believe drive the intuition that the Man does not understand the language he computationally simulates. Consideration of this version might lend greater reliability to judgments about whether the Man understands the language he simulates. It is to be expected that we are better at telling the difference between a person who understands and one who does not understand a language we ourselves understand, than at telling the difference between a person who does not understand and a person who does understand a language we ourselves do not understand. The English Room Argument eliminates the possibility of basing the judgment that the Man does not understand Chinese on our own condition. Because most people in the debate over the Chinese Room Argument do not understand Chinese, to most people in this debate, Chinese symbols look meaningless. In fact, this is why Searle has substituted them for combinations of SQUIGGLE SQUOGGLES. But the fact that Chinese symbols look meaningless now, without running the program for understanding Chinese, does not imply that under the conditions supposed by the Chinese Room Argument, we would not understand. We might very well change our view that we do not understand Chinese were we to be running the program. This would not be more surprising than the fact that if some movements of molecules in our brains were changed, we would understand languages we

don't currently understand and we would have conscious experiences we don't currently enjoy.

The Man Can Disambiguate Symbols that Look the Same

A key feature of understanding the semantics of a language is the ability to disambiguate polysemous symbols. Can the Chinese Man do this with English symbols? Let us look at some disambiguation tasks. The symbol for financial banks and for the sides of water systems has the same shape: "banks." One way to tell whether the Chinese Man can disambiguate the symbol is to see whether he reliably uses the symbol appropriately. If the Man does not, then this lends credence to the idea that he does not understand, and if he does, then this supports the idea that he does understand English.

Suppose the Chinese Man received the set of symbols "Let's go sailing at the bank." A genuine English speaker would answer "Cool, let's do it" or "I don't want to get wet in that cold water," for example. So would the Chinese Man. Like the genuine English speaker, the Chinese Man in the English Room Argument would not output symbols like "That's crazy. The bank guards will kick you out, thinking you are a menace to the safety of the clients."

The Chinese Man is supposed to be behaviorally indistinguishable from a real English speaker. Minimally, normal persons would exhibit appropriate verbal behavior requiring the disambiguation of polysemous symbols like the present one. One could also ask the Chinese Man directly, "What is the meaning of 'bank'?" This question would be on a par with any other normal question that could be put to him, which would be answered as an authentic speaker of English would. Given that he is behaviorally equivalent to a true speaker of English and that his rule-book is complete, he could answer something like this: "There are two meanings of 'bank.' One refers to the sides of water systems and the other is about institutions where people keep money and take out loans."

On the flip-side, suppose someone asked the Chinese Man the same question in Chinese (maintaining the "bank" symbol in English), his otherwise only language. We can expect him to answer appropriately. In line with Searle's design, what would enable him to say this is that the rule-book would contain a complex set of commands such that if asked for the meaning of "bank" he would be guided to say things like the mentioned response example. Further, understanding the semantics of symbols provides us with the ability to make correct inferences. Take, for example, the following inference:

1. All banks are financial institutions.
2. All banks are along the edges of waterways.
3. Therefore, all banks are financial institutions along the edges of waterways.

We all agree that the premises are true and the logical syntax seems to be validly applied. But this, however, by no means convinces anyone committed to the premises that they are committed to the conclusion. Rather, it is easy to respond that the word "bank" is being used in different senses and that consequently, the truth of the premises, under the plausibly true interpretation, does not logically imply the conclusion. The Chinese Man would respond to such an argument in an equivalent way.

Consider another example. Read closely the story of *The Wrestler*:

> Rocky slowly got up from the mat, planning his escape. He hesitated a moment and thought. Things were not going well. What bothered him most was being held, especially since the charge against him had been weak. He considered his present situation. The lock that held him was strong but he thought he could break it. (Anderson, Reynolds, Schallert, and Goetz, 1977, p. 372)

The passage seems clear. In understanding it, certain computational channels are activated in readers, and certain inferences can be drawn from it which enable readers to answer questions. Like other readers, the Chinese person would be able to answer the question of who the wrestler in the story is: Rocky.

But now, answer the following question with respect to the text: "Who is the inmate?" Your answer is probably something like this: "There are no mentioned inmates in the text — there is a wrestler, not an inmate!" The Chinese Man would answer the same. Now read the passage again with close understanding and think of it as a description of a prison escape, titling it *The Prisoner*. Here too, there is a clear and distinct meaning, which will be associated with the activation of another computational passage and other behaviors. Now you can answer the question above: "Rocky is the inmate." The Chinese Man would exhibit the same response patterns.

Intuitively, without much theory in place, the Chinese Man's competence at respecting the semantic boundaries and conditions of application of symbols that look the same is some evidence that the Man really understands the semantics of the text at issue. Now, take the structure building framework theory of understanding (Gernsbarcher, Varner, and Faust, 1990). Perhaps this specific psychological theory is correct, and perhaps not. Let us, for a moment, suppose that it is right, as a proxy for whichever is truly correct. The structure building framework theory says that we understand a text by establishing a frame or structure onto which new information is mapped. If we do not map the information, we build a new structure onto which further information can be mapped. The behavior of an unquestionably genuine English speaker (you, in this case) would be explained by structure building framework theory: the two instances of reading the text about Rocky had titles which established distinct general subject-matters or "structures," onto which the rest of the narrative was mapped. Before reading the passage under the title of *The Prisoner*, the subject relevant to the subsequent

question was not aptly fixed, and thus, the reader did not know what prisoner was being inquired about. The structure building framework theory could explain the pattern of your behavior; and since the Chinese Man displays the same pattern of behavior, there is a prima facie case that he understands English through similar mechanisms.

Whichever theory of understanding truly explains your lexical disambiguation behavior patterns, it will posit mental properties which would seem to similarly explain the Chinese Man's behavior patterns. Thus, given that your understanding explains your behavior, the Chinese Man's understanding would similarly explain the Chinese Man's behavior.

The Symbols Seem to Elicit Comprehension-Mediated Responses

A key feature of the human mind is that it has a set of concepts with semantic contents, whose relations help humans remember, perceive, and orient themselves in their environment. The degree of strength of the semantic relations between concepts varies. Semantic priming is a mechanism hypothesized to be one by which this web of ideas operates (Neely, 1976; Whitney, 1998, pp. 92–93). The basic mechanisms posited by the theory of semantic priming are the facilitation and the inhibition of the processing of incoming information depending on their semantic ties. For example, in a lexical decision task, a person decides as fast as she can whether a second string of letters is a word. The person might be shown the string ROBIN or XXXX, and then BIRD. BIRD is obviously more semantically connected to ROBIN than to XXXX. Persons are consistently faster at deciding whether BIRD is a word when the first shown string is ROBIN than when the first string is XXXX. There is a facilitation effect from ROBIN to BIRD; the person is semantically primed for BIRD. Glenberg (1997) suggests that semantic priming is an instance of the operation of a mesh, where a mesh is the encoding of possible interactions between our embodied selves and the rest of the world. Further, in Glenberg's scheme, a person's understanding of a lexical item crucially involves the person's encoding of a set of actions that might be taken in response to the reference of meaningful fragments. Glenberg believes that such a conception erases some of the experimental embarrassments of the theory of semantic priming. In particular, semantic priming theory is challenged by evidence showing that priming may be due to non-permanent links, as early semantic priming theories denied (Glenberg 1997, p. 14 cites McKoon and Ratcliff, 1986) and that priming is observed between strings which are not related by their presumed semantics (Glenberg, 1997, p. 14 cites Shelton and Martin, 1992).

Nevertheless, whether semantic priming is true as originally envisaged or as envisaged by mesh theory, it still involves mental structures which have semantic content in that it involves mental structures *about* the way the world is, and it is the relations between items in those mental structures which drive the observed

priming effects. With this in mind, the operation of a mesh is itself the operation of a semantic relation — that is, the operation of a relation between elements with intentionality in a mental structure. Glenberg simply has a particular conception of the semantics of such mental structures, with an emphasis on perception and action.[3] Further, the effects of semantic priming are in response to observed strings of lexical items. People who understand a language need to have appropriate links between their lexical items and their mental structure in order to generate the observed responses.

The response times the Chinese Man would display would be indistinguishable from those of an authentic English speaker. In fact, all observations relevantly made in the semantic priming literature and the mesh literature will be observed in the Chinese Man, because as Searle supposes, the two are behaviorally identical. The hypothesis that these cases are explained by the idea that the Chinese Man, like other humans, has a mental structure representing the world, including himself, that mediates between his different observations and responses, should not be underestimated. The degree of detailed symmetry in expected responses between the Chinese Man and those of a genuine English speaker is impressive.

The Chinese Man Seems to Form an Intentional Mental Representation

A core feature of a person who understands a meaningful declarative text is that she forms an intentional mental representation of what the world would be like if the text were true. There is a consensus among psychologists that there are various stages in text processing, each stage dealing with a distinctive aspect (Whitney, 1998, p. 259), and at some point in the processing of a text, the reader forms a mental representation of what the sentence is about. This is a common understanding amongst neutral, nativist, and non-nativist views of language. Thus, truth-conditional semantics (e.g., Frege, 1879), conceptual semantics (e.g., Jackendoff, 1992), and cognitive grammar theory (e.g., Lakoff, 1987) accept this. It is also accepted by minimalist and constructivist theories (e.g., Kintsch, 1988, 1992; McKoon and Ratcliff, 1992). Minimalism theorizes that understanding a text involves conceptual connections with structures close to the form of the text itself and involves minimal connections to information from areas of the mind which are not specifically linguistic. The constructivist view takes it that the use of areas not specific to language is an important part of understanding. The mental representation studied within these theories would

[3]In psychology, "semantic" and "episodic memory" are typically distinguished. The intended use of "semantic" here primarily involves the philosopher's notion that an item of memory has semantics if it has referential meaning. This philosophical notion is akin to the psychologist's declarative memory.

seem to be the one in which Searle (1980) is interested. Whichever theories in this domain turn out to be true, the Chinese Man would provide equal experimental evidence for them as any English-speaking person would. Here is one example of the kind of evidence that might be put forth:

> Roger broke up with his long-time girlfriend and moved in with Gail. The ex-girlfriend is quite upset and is trying to win Roger back and convince him to leave Gail. He'd left her for Gail before, she thought, and he might leave Gail for her now.
>
> Roger was invited to a party with old friends, and was happy to see Haruna, Elizabeth, Emily, Matt and Sagar. Gail, however, could not go to the party as she was feeling sick, so she stayed home. Roger, on the other hand, having a better time than expected, stayed until late. When he arrived home, Gail was distressed and asked "Tell me the truth. Was *she* there?" (adapted from Whitney, 1998, pp. 245–246)

Now, in Gail's question, who is *she*? You and everyone who understands the passage knows who it is: the ex-girlfriend. The Chinese Man would respond just as you would. This is quite surprising if he does not understand what is going on.

The Man in the Chinese Room Argument Seems to Know the Reference of the Symbols

Let us consider the original Chinese Room Argument again. Experimenters might want to determine whether the Man has the capacity to match Chinese symbols to the things to which they refer. If he could, this would support the claim that the Man is picking up the referential meaning of the Chinese language. Thus, they might ask the Man, in English and in Chinese on different occasions, to identify a 笔 (which means PENCIL). The experimenters would find that in both cases the Man will identify a pencil, for example, by pointing at one or by saying in Chinese, or English if required, that it is the object in his hand with which he writes. The Chinese Room Argument asks us to suppose that the Man is behaviorally equivalent to a man who understands Chinese. Thus, when the Man is asked in English, he will look in his rule-book and find a command taking him from his wish to identify a 笔, to the identification of a pencil. Commands will enable him to answer appropriately.

Someone outside the Room might hand the Man a pencil and ask him to say what it is in Chinese. Some might be surprised to know that he would answer correctly, with the 笔 symbol. How can this be? The answer is that if the rule-book is complete, the Man's behavior in response to the question will be based on a command functionally equivalent to "If asked to name a pencil in Chinese, output 笔." This is so whether the question is asked in English or Chinese. The Chinese Room Argument already supposes that while the Man computes and

behaves as a person who understands Chinese, he speaks English, for this is the language in which the rule-book is written.

Another evidentially relevant case is one in which experimenters tested the Man's ability to apply color terms correctly. The conscious experiences associated with colors are a key aspect of our understanding of color terms. The conscious experience associated with each color determines the semantics of our color terms and are a key aspect of what enables us to apply them correctly. Does the Man apply color terms correctly?

Let us focus on the English Room Argument version to look at this potential experiment. Experimenters could ask the Chinese Man (in English or Chinese) to identify green, orange, magenta, red, and black items in the Room. The Chinese Man is behaviorally equivalent to a man who really understands English. A man who genuinely understands English would, under normal circumstances, be able to identify green, orange, magenta, red, and black objects. The Chinese Man would be enabled to find and differentiate the differently colored objects by a command in the rule-book which would instruct him, like in the PENCIL case, to identify objects of the desired color. For example, an available command would be functionally equivalent to: "if asked to identify 绿色, identify green." Given that he can identify green things just as he can visually identify the symbols, this should not be a problem.

Naturally, this surprising ability of the Chinese Man in the English Room Argument lends credence to the hypothesis that he understands English. The Chinese Man consistently applies the term "green" to green items; as with the rest of the terms and the colors to which they refer. Further, the Man does this in quite a flexible and fine-grained manner. Consider Heider's (1972) classic refutation of a strong version of the Sapir–Whorf hypothesis (Hunt and Agnoli, 1991). According to this version of the theory, we cannot cognize outside the basic categories of our language. Heider studied the Dani tribe in New Guinea, which has only two basic terms to name colors. Heider showed people of the Dani tribe chips of one color (say, focal red). Sometime later she showed them chips of many colors, including a shade of the previously shown color (say, an off-shade red), and asked them what color she had showed them before. The Dani, like English speakers, remembered focal colors better than off-shade colors, and distinguished between colors for which there was only one basic term. The Dani understood that each of their terms could refer to various kinds of colors, even if they were grouped together under one term.

Similarly, the Chinese Man might be shown a focal red chip at one time, and subsequently be asked, in English and in Chinese on different occasions, to identify chips with the same color. Like the Dani, the Chinese Man would reliably identify the chips with the correct color. This suggests that the Man understands that there are various colors to which "red" applies and knows which colors those

are. This is an important part of understanding the meaning of "red." The task normally requires appropriate interactions between a person's perceptual, linguistic, and mnemonic systems, which is part of what enables people who understand a language to understand it.

A "Psychological Version" of the Turing Test

The Turing test is a recognized high-standard test designed to detect the existence of a target entity's intelligence. Originally, it was meant to be applied to an artificial machine, and involved the idea that if a non-expert person could not tell the difference between when she talked with the machine or with a human being, then the artificial machine would have provided sufficient evidence to conclude that it thinks. In the Chinese Room Argument, the Man takes the place of the artificial machine. The Man is behaviorally indistinguishable from a man who genuinely speaks Chinese. Thus, the Man passes a highly demanding test for understanding Chinese, one which no artificial computer has ever passed.

As Copeland (2004, p. 435) notes, there is evidence that Turing did not intend to *define* thinking in terms of the ability to pass the imitation game, which has come to be known as the Turing test. Rather, Turing intended to set it up as a sufficient condition. However, it is not true that no textual evidence exists that Turing intended to give a behaviorist definition of thinking since Turing explicitly stated that he wished to *replace* the question of the possibility of a thinking machine with the question of the machine's ability to be behaviorally indistinguishable from a person questioned by an interrogator under the conditions of the Turing test (Turing, 1950/2004, p. 441; cf. Copeland, 2004, p. 435). Out of a general scepticism about the ability to speak meaningfully about such inner "unobservable" states, behaviorist definitions of mentality were in vogue in Turing's day (see for instance, Hempel, 1949). Turing also claimed that the question, "Can machines think?" would be too susceptible to be analyzed and answered by a Gallup poll, which he considered absurd. However, this argument is like saying that the question of whether the Earth orbits around the Sun should be replaced by the question of whether an observer would see the position of the Sun as would be expected if the Earth orbited around it, because the former question would be answered by a Gallup poll, which would be absurd. The argument as a whole is not convincing.

The correct form of the argument is, in my view, that a machine's ability to pass the Turing test constitutes significant evidence that the machine thinks, just as an observer's seeing the systematic position of the Sun from the Earth as if the Earth orbited around it is strong evidence that the Earth indeed orbits around the Sun — Gallup polls and replacements of questions being irrelevant.

Let us, however, raise the bar; let us test whether the Man passes more exacting versions of the Turing test next.[4] Unlike what is allowed in the original Turing test, suppose a psychologist took the role of the interrogator. She might want to try distinguishing the Man by adding experimental conditions used to identify the ways in which a person who truly understands does in fact understand. For example, she might perform Chinese versions of the tests relevant to language understanding of the kind we considered above. The entity that performs outside the norm would be identified as the Man, and consequently, the Man would not pass the psychological version of the Turing test.

The problem is that the Man is, by hypothesis, behaviorally equivalent to a real Chinese speaker, and consequently would pass this version of the Turing test. The Man would display in minute detail the behavioral grounds which would justify the attribution of Chinese understanding to genuine Chinese-understanding humans. Timing provides insight into various aspects of the processing characteristic of understanding and is consequently used in various psychological experiments. Here also, the Man would display the same timing as someone who was processing in the way human understanding exhibits (cf. Dennett, 1987).

French (1990), however, thinks an artificial computer could not show such timings. Considering the lexical decision task used to detect semantic priming, for concreteness, he states:

> The machine would invariably fail this type of test because there is no a priori way of determining associative strengths . . . . Virtually the only way a machine could determine, even on average, all of the associative strengths between human concepts is to have experienced the world as the human candidate and the interviewers had. (cited in Copeland, 2000a, p. 535)

Mutatis mutandis, it might be argued that the Man could not display those timings either. But the argument is weak. For one, the associative strengths are part of

---

[4]Notice, however, that the issues of thinking and relevantly, of understanding, are still in question. To try to deflect the force of the Chinese Room Argument by saying that Searle is "merely exploring facts about the English word *understand*" (Pinker, 1997, p. 95) is as convincing as being told that finding a place where there is no light but there is electromagnetism does not refute the electromagnetic theory of light because we are just exploring the stereotypical meaning of the word *light*. There are four possible explanations of such a scenario compatible with the truth of a version of the electromagnetic theory of light: (i) to the contrary, that in such a putative case there is light, but it is just that there is not enough or of the right frequency for us to see; or (ii) that, in fact, there is no electromagnetism in the supposed place; or (iii) that "light" refers to the visible portion of electromagnetic waves, and that consequently, electromagnetic waves of the different kind present in the case at issue are not of the relevant kind; or (iv) a combination of the above. Otherwise, the electromagnetic theory of light would be refuted (compare Churchland and Churchland, 1990; Turing, 1950/2004; Turing, Braithwaite, Jefferson, and Newman, 1952). By hypothesis, in the Chinese Room Argument there are the right kinds of computations implemented, so if there is no understanding, the computational theory of mind is refuted. Pinker simply fails to address Searle's challenge.

the computational dynamics a Chinese speaker displays. On the supposition that a machine could mimic the computational processes of the Chinese speaker, the maker of the machine would link various symbols to other internal processes functionally equivalent to their mental semantics, and not to others. French proposes his "limit" on the Turing test on the basis of the idea that semantic associative strength could not be determined by something other than having had certain human experiences, which by hypothesis the machine has not had. But by that reasoning, no machine could ever pass the original (without an expert interrogator) version of the Turing test, since the machine would have to display behavior normally determined by having had certain human experiences, which by hypothesis, the machine has not had. By French's standards, the fact that a machine has not had the human experiences which normally enable a human to think and understand certain things prevents the machine from making associative semantic links. By this standard, the machine could not even generate a credible answer to the relatively undemanding question "How was your morning coffee?" because it has never had a morning coffee, which must, in my view, be wrong. A psychologist interrogator in the Turing test would not reliably be able to detect the Chinese Man. Consequently, the Man would pass a very demanding Turing test for understanding Chinese.

The "Mathematical Version" of the Turing Test

Penrose (1990, 1994) has been a consistent critic of computational cognitive science precisely because Gödel's and others' results indicate that the outputs of mathematicians are different from the outputs of Turing machines, and that consequently, people are not computers. Suppose Penrose spoke Chinese and was the expert interrogator. He might inquire into whether the people in the other rooms could arrive at uncomputable results, using Chinese symbols. The person who cannot is the Man. He might say to them in Chinese, "tell me whether *pi* has a last decimal." Both respondents would answer the same: "否" (No). Most people have not worked this out in detail, but take it on authority that this is so, or use mathematical induction after a sufficiently high, but finite and Turing-computable, number of checks. Similarly, the Man would have encoded or processed equivalent information. Why should we have a higher standard for the Man?

On the other hand, suppose Penrose was the interrogator and Gödel was the other person in one of the rooms. Gödel is well-known for proving that there are some mathematical truths which cannot be proved by a finite set of rules. Turing's (1936/2004) work itself involved this claim, and addressed it as the Mathematical Objection (Piccinini, 2003; Turing, 1950/2004). Perhaps Penrose would be able to differentiate (a Chinese-speaking) Gödel from the Man. The standard of such a test was pointed out by Max Newman in his discussions with

Turing and others (Turing, Braithwaite, Jefferson, and Newman, 1952, p. 505). If we add Newman's standard to that of the Turing test, we get the Mathematical Version of the Turing test.

But the fact that Penrose would be able to differentiate the Man from Gödel should prove nothing. First, Penrose would be able to differentiate Gödel from any probable genuine Chinese speaker, given Gödel's exceptional mathematical sophistication. By the considered standards, Penrose would only let a handful of people, if any, of the Chinese population pass the test for understanding Chinese. Secondly, as Copeland has noted, it might well be that a human computer is not equivalent to a Turing machine, but to another sort of computing machine Turing designed in his Ph.D. thesis: an O-machine (Turing, 1939; see Copeland, 1998, 2000b). Supposing this is so, if Penrose could design a reliable method for weeding out Turing machine-equivalent persons, and keeping O-machine-equivalent people, then perhaps this could hamper the Man's ability to pass this version of the Turing test. However, the operation of the Man in accordance with the rule-book might well be such that it allows him to instantiate an O-machine. To introduce a random element characteristic of O-machines (Copeland, 1998), the rule-book may contain an instruction of the form "If φ, then take symbols from basket 33 and 249, copy them, shuffle the copies, put them in basket 841 and pick any out." Simulating a Chinese mathematician, the Man could conceivably return the values for Turing's Halting function H (x,y) (Copeland, 1998, p. 130; Turing, 1936/2004).

The "Interactive Version" of the Turing Test

In the Interactive Version of the Turing test, in addition to text, the interrogator and the entities he tries to distinguish can exchange other things. For example, the interrogator may introduce a beetle, pass it to the Man and ask what it is, in Chinese. As with the case of PENCIL above, provided the Man knows that it is a beetle, the Man will be able to follow a rule which will provide him with adequate Chinese labels and descriptions as answers to questions, which will make him behaviorally indistinguishable from a Chinese speaker who also receives the beetle.

Now, suppose the interrogator introduced some unspecified objects in three boxes, kept one and handed each of the other two boxes to the other two people in the Turing test, and told them in Chinese that they will call the thing in the box a "甲蟲" (BEETLE) [cf. Wittgenstein, 1953]. Notice that if they had different kinds of objects, they could figure this out by a series of answers and questions in Chinese about whether the object is biological, its color, the number of legs it has, what it eats, whether it flies, and so forth. The Man would display a competence for communicating in Chinese and exhibit behavior equivalent to that characteristic of coming to know that he has the same kind or a different

kind of object in the box as the interrogator, despite the label. The Man would pass this very stringent version of the Turing test. This is surprising if he does not understand.

The Persons Reply

The first central thesis of the persons reply is that the Man is the entity implementing the computational properties characteristic of understanding Chinese, and that consequently, the Man is the Computer whose mental properties are relevant to the testing of computational cognitive science (Restrepo, 2012a). Exploring the considerations that warrant this judgment and the apparent relative weakness of the rationales of the opposing view, the persons reply is in sharp contrast to the logical (Copeland, 2002) and systems replies (Block, 1980, 1995, 2002). Although Searle does not use the same arguments as the persons reply, we agree on this. We also agree that there is a real and important question of whether the Man understands Chinese (cf. Pinker, 1997).

Scientific realism implies that scientific inquiry is a reliable truth-seeking operation. The methods of science are particularly good at telling us about nature and its causal structure. Knowledge of the causal powers of different types of things differentiated by the sciences is what enables us to increasingly explain and predict nature, as well as to create new technology. Psychology has its own way of sorting elements in nature, which focuses on the properties and causal powers of the things in which it is interested: mental things. This involves the claim that some properties of things are relevant to psychology and some are not. For example, thinking is an interesting property for psychology, and to it there corresponds a characteristic set of causal powers. The brain might have the consistency of cold porridge, but in trying to duplicate the properties and causal powers of the brain which are relevant to psychology, duplicating the consistency of cold porridge is misguided (see also Restrepo, 2012b; Turing et al., 1952, p. 495). It is thinking that we are interested in.

There is an important class of properties and causal powers which are hard to directly observe and interpret. Having a negative charge, being radioactive, various aspects of other people's and even our own mental lives are not directly epistemically accessible. However, we do have reason to believe various things have a negative charge, are radioactive, other people have thoughts, and we have various psychological properties we are aware of and others we are not. A particularly good way of discovering these things is through experimental methods, which help filter out the properties we are not interested in and provide us with results that can give us reason to know the character of those properties and causal powers of interest. In the case of the Man, the experimentally selected external appearance Searle demeans is important because it gives us evidence of the internal structure which gives rise to it. The best

explanation for the explanatory, predictive, and technological success of well confirmed scientific theories is that they are (approximately) true. The theory that the Man understands Chinese receives robust confirmation from all of the psychological and artificial intelligence tests performed on him, and we are consequently warranted in believing its truth, which is the second thesis of the persons reply. Given that the theory of computational cognitive science successfully overcomes the experimental risk represented by the Chinese Room scenario, this foundational theory itself receives confirmation from the proposed proper understanding of the Chinese Room Argument.

References

Abelson, R. (1980). Searle's argument is just a set of Chinese symbols. *Behavioral and Brain Sciences, 3,* 424–425.

Anderson, C., Reynolds, R., Schallert, D., and Goetz, E. (1977). Frameworks for comprehending discourse. *American Education Research Journal, 14,* 367–381.

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, L. Gleitman, S. Kosslyn, S. Smith, and S. Sternberg (Eds.), *An invitation to cognitive science* (pp. 377–426). Cambridge, Massachusetts: MIT Press. (Retrieved on Aug. 22, 2010, from http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/msb.html)

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.

Churchland, P., and Churchland, P. (1990). Could a machine think? *Scientific American, 262,* 32–37.

Copeland, J. (1998). Turing's O-machines, Searle, Penrose, and the brain. *Analysis, 58,* 128–138.

Copeland, J. (2000a). The Turing test. *Minds and Machines, 10,* 519–539.

Copeland, J. (2000b). Narrow versus wide mechanism: Including a re-examination of Turing's views on the mind–machine issue. *Journal of Philosophy, 97,* 5–32.

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). Introduction to *Computing Machinery and Intelligence*. In J. Copeland (Ed.), *The essential Turing* (pp. 433–440). Oxford: Oxford University Press.

Dennett, D. (1987). Fast thinking. In *The intentional stance* (pp. 324–337). Cambridge, Massachusetts: MIT Press.

Dix, A. (2005). The brain and the web: A quick backup in case of accidents. *Interfaces, 65,* 6–7.

Frege, G. (1879). *Conceptual notation and related articles* [T.W. Bynum, Trans.]. Oxford: Clarendon Press.

French, R. (1990). Subcognition and the limits of the Turing test. *Mind, 99,* 53–65.

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.

Glenberg, A. (1997). What memory is for. *Behavioral and Brain Sciences, 20,* 1–55.

Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology, 93,* 10–20.

Hempel, C. (1949). The logical analysis of psychology. In H. Feigl and W. Sellars (Eds.), *Readings in philosophical analysis* (pp. 373–384). New York: Appleton–Century–Crofts.

Hunt, E., and Agnoli, F. (1991). The Whorfian hypothesis: A cognitive psychology perspective. *Psychological Review, 98,* 377–389.

Jackendoff, R. (1992). *Languages of the mind: Essays on mental representation*. Cambridge, Massachusetts: MIT Press.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review, 95,* 163–182.
Kintsch, W. (1992). A cognitive architecture for comprehension. In H.L. Pick, Jr., P.W. Van den Broek, and D.C. Knill (Eds.), *Cognition: Conceptual and methodological issues*. Washington, DC: American Psychological Association.
Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.
McKoon, G., and Ratcliff, R. (1986). Automatic activation of episodic information in a semantic memory task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 1–15.
McKoon, G., and Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99,* 440–466.
Neely, J. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory and Cognition, 4,* 648–654.
Penrose, R. (1990). Précis of *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. *Behavioral and Brain Sciences, 13,* 643–655, 692–705.
Penrose, R. (1994). *Shadows of the mind: A search for the missing science of consciousness*. Oxford: Oxford University Press.
Piccinini, G. (2003). Alan Turing and the Mathematical Objection. *Minds and Machines, 13*, 23–48.
Pinker, S. (1997). *How the mind works*. New York: W.W. Norton.
Restrepo, R. (2009). Russell's structuralism and the supposed death of computational cognitive science. *Minds and Machines, 19,* 181–197.
Restrepo, R. (2012a). Computers, persons, and the Chinese Room. Part 1: The human computer. *Journal of Mind and Behavior, 33,* 27–48.
Restrepo, R. (2012b). Multiple realizability and Novel causal powers. *Abstracta, 6,* 216–230.
Sandberg, A., and Bostrom, N. (2008). *Whole brain emulation: A roadmap*. Technical Report #2008–3, Future of Humanity Institute, Oxford University (Retrieved on Nov. 23, 2011 from http://www.fhi.ox.ac.uk/Reports/2008–3.pdf)
Shelton, J. R., and Martin, R. C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1191–1210.
Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3,* 417–457.
Searle, J. (2002). Twenty-one years in the Chinese Room. In J. Preston and M. Bishop (Eds.), *Views into the Chinese Room: New essays on Searle and artificial intelligence* (pp. 51–69). Oxford: Oxford University Press.
Turing, A. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society, 45,* 161–228.
Turing, A. (2004). On computable numbers, with an application to the *Entscheidungsproblem*. In J. Copeland (Ed.), *The essential Turing* (pp. 58–96). Oxford: Oxford University Press. (Originally published in 1936)
Turing, A. (2004). Computing machinery and intelligence. In J. Copeland (Ed.), *The essential Turing* (pp. 441–464). Oxford: Oxford University Press. (Originally published in 1950)
Turing, A., Braithwaite, R., Jefferson, G., and Newman, M. (1952). Can automatic calculating machines be said to think? In J. Copeland (Ed.), *The essential Turing* (pp. 494–506). Oxford: Oxford University Press. [Broadcast by BBC Radio on January 14, 1952]
Whitney, P. (1998). *The psychology of language*. New York: Houghton Mifflin.
Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell Publishing.